The Effect of Different Observational Session Lengths on the Representativeness of Behavioural Data

Oliver Charles Mudford

University of Canterbury, Christchurch

A thesis submitted in partial fulfillment

of the requirements for the degree of

MASTER OF ARTS IN PSYCHOLOGY

February, 1987

Acknowledgements

I offer many thanks to Dr. Nirbhay Singh (Department of Psychology, University of Canterbury) and Associate Professor Ivan Beale (Department of Psychology, University of Auckland) who supervised this thesis. I am grateful for the assistance and cooperation from residents and staff, particularly Ms. Jane Penney, at Templeton Hospital and Training School, Christchurch.

A preliminary report on Study 2 was presented as a conference paper to the symposium of the Division of Behaviour Analysis at the 1986 (August) Conference of the New Zealand Psychological Society at Dunedin.

Contents

Abstract

The representativeness of behavioural observation samples with durations of less than the whole time of interest was investigated. A real-time recording system was developed to quantify the performance of five profoundly mentally retarded physically handicapped adult students in an institutional training setting. Performance was measured on six mutually exclusive and exhaustive behaviour categories throughout 2.5-hour morning and afternoon training sessions with each subject. Passive behaviour, i.e., doing nothing, was the most predominant category (mean = 46% of session). Sample observation sessions with durations ranging from 15 to 135 minutes were computer-simulated from the whole session (150 minute) records. It was found that the representativeness of these samples, when compared to the whole session, was a function of the relative duration of the behavioural categories and of sample duration. The occurrence of high relative duration behaviours (>50% of a session) was estimated to within 20% error by samples of less than 60-minute but low relative duration behaviours (1-3%) were inadequately quantified even from 135-minute samples. Implications of the findings for behaviour analysts were discussed with the recommendation that the adequacy of observational session duration should routinely be empirically determined.

The Effect of Different Observational Session Lengths on the Representativeness of Behavioural Data

General Introduction

Direct observation and measurement of behaviour for assessment of human performance is a hallmark of applied behaviour analysis methodology. This can be contrasted with more traditional attribution of traits and application of global rating scales of impressions of behaviour (Kazdin, 1984a). With mentally retarded persons (as well as others) valid assessment procedures are required for evaluation of services (e.g., Alevizos, DeRisi, Liberman, Eckman, & Callahan, 1978; Green et al., 1986; Repp & Barton, 1980; Van Biervliet, 1982) and of therapeutic interventions (e.g., Burgio, Page, & Capriotti, 1985; Landesman-Dwyer & Sackett, 1978). "We suggest that direct observation under everyday living conditions is the only measurement method that can yield valid information concerning the ethical, political, and economic decisions made by governmental and other institutions that affect the lives of retarded people" (Sackett & Landesman-Dwyer, 1977, p. 28).

Ever since the first formal description of the observation methods of applied behaviour analysis (Bijou, Peterson, & Ault, 1968), these methods have been critically examined both empirically and theoretically. The issue of the validity of the procedures employed to acquire data for quantification of behaviour has become of increasing concern. We need valid

information (Sackett & Landesman-Dwyer, 1977) but we can only assess its degree of validity through evaluation of the methods employed to acquire it.

A useful model for describing the requirements of a valid observational system is Generalizability (G) Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Intuitively, 'validity' can be seen as meaning correspondence with the true state-of-affairs, e.g., how well the data obtained by a particular observational procedure reflect the true frequencies and/or durations of the behaviours of interest. G-theory views validity as correspondence between obtained data and a universe of data which could be obtained from all possible assessments. Cronbach et al. (1972) point out that their concept of 'universes' is logically and conceptually identical to that of the more familiar concept of 'populations', with the latter being reserved for subjects in G-theory.

There has been argument both for (Strossen, Coates, & Thoresen, 1979) and against (Jones, 1977) the application of the methods prescribed by G-theory to assess the validity of the intensive and repeated direct observation procedures commonly employed in applied behaviour analysis. Violation of the assumptions in the analysis of variance (ANOVA) models employed in G-theory have been identified as being due to non-independence of successive samples of behaviour (auto-correlation) and, in intervention studies, the lack of a steady-state (unless the intervention does not affect behaviour).

The six universes of interest to behavioural assessors (A-F

below) have been identified by Cone (1977). Although he did not extend his analysis of the utility of G-theory to the type of intensive assessment procedures examined in the present work, a case can be made to include behavioural analysts' concerns about validity under the same six headings by inclusion of examples. Further examples may be found in Foster and Cone's (1980) review of direct observation which takes a G-theory perspective.

A) The universe of Scorers refers to generalizability across all possible observers. The procedures developed for assessing inter-observer reliability (House, House, & Campbell, 1981; Maclean, Tapp, & Johnson, 1985) have addressed this universe more or less adequately (Berk, 1979; Hollenbeck, 1978; Jones, 1977). Usually the ANOVA models required by G-theory have not been employed in estimating reliability although there have been exceptions (Berk, 1979; Booth, Mitchell, & Solin, 1979; Jones, Reid, & Patterson, 1975). A point to be made here is that, in the assessment of interobserver agreement, it appears acceptable to employ the concepts of G-theory without adopting the methods. That is, adopting the spirit of the theory but not its application to the letter. Foster and Cone (1980) and Jones (1977), for example, call on G-theory in a conceptual manner rather than adopting the full methodology suggested by Cronbach et al. (1972).

B) Items refers to correlation between behaviours observed which are in the same class. When assessing the behaviours of clinically referred 'problem' boys, Jones et al. (1975) found that the operationally defined items grouped as aversive,

hostile, and irritating behaviours could be combined to form a class of behaviour labelled Total Deviant without loss of validity of the assessment instrument, the Behavioral Coding System.

C) Times refers to generalization across occasions of measurement." ... the question is to what extent data obtained on one occasion of measurement are comparable to those obtained from other samples of the entire universe of measurement occasions or times. A slightly different temporal stability question has to do with the extent to which scores may vary over time intervals differing in length" (Cone, 1977, p.418). An example of generalizability in the universe of Times is the finding of Alevizos et al. (1978) that 5 seconds observation of psychiatric patients in an institutional setting produced data not significantly different from those obtained from 30-second observations.

D) Settings refers to the generalizability of data across the situations in which the behaviours of interest may occur. This universe is addressed when generalization across settings has been assessed in intervention studies employing the multiple baseline across settings experimental design (e.g., Landesman-Dwyer & Sackett, 1978; Odom, Hoyson, Jamieson, & Strain, 1985).

E) Methods refers to the comparability of data obtained by different measurement procedures. Cone (1977) has discussed the importance of this universe for behavioural assessors in terms of the comparability of, for example, self-report versus observational data. For behavioural analysts the Methods universe

can be seen to include comparison of data obtained by different observational procedures (e.g., Green, McCoy, Burns, & Smith, 1982; Towns, Singh, & Beale, 1984) and technologies (e.g., Linscheid, Feiner, & Sostek, 1984).

F) Dimensions refers to the degree of relationship between different classes of observed behaviour. The multiple baseline across behaviours experimental design used in clinical studies can be viewed as relevent to this universe (e.g., Buell, Stoddard, Harris, & Baer, 1968). The concept of behavioural interrelationship emphasises the importance of the Dimensions universe to behaviour analysts (Voeltz & Evans, 1982).

In this thesis a facet of the generalizability universe of Times was investigated. The validity of data collected from observational sessions of different durations was assessed.

Although the validity of behavioural data needs to be evaluated by behaviour therapists involved in intervention studies, the emphasis in the present series of experiments was on elucidating questions about the validity of data for assessors of services and consumers of their reports. Study 1 was designed to describe the performance of physically handicapped profoundly mentally retarded adult students of an institutional training setting. As with all observational studies, Study 1 potentially contributes to knowledge of the lifestyles of mentally retarded people. The main purpose, however, was to describe the database sampled in Study 2.

Study 2 sought to assess the generalizability of data

obtained from samples which were not exhaustive, i.e., samples with durations less than the whole time of interest (a training session). The whole session was the universe in Times for which generalizability was investigated. This represents an attempt to find a time-economical method of assessment of students' performance within a training session while, as far as possible, preserving the validity obtained from exhaustive sampling.

Study 1

Quantified description of the behaviours of physically handicapped profoundly mentally retarded adults in an institutional training setting

Quantification of performance

To adequately quantify observed subjects' behaviours several basic parameters always need to be considered: frequency, duration, and inter-response time (IRT) (Altmann, 1974). Frequency data are concerned with how many times a behaviour occurs. Duration data tell us how long a behaviour persists once initiated and, in the case of relative duration, how much of an observational session is taken up with a behaviour. IRT data give information as to the amount of time passing between instances of behaviour.

Sometimes other dimensions of behaviour are quantified, e.g., intensity (the strength of a response) and topography (the observed form of the response). In most studies these other dimensions are subsumed, either explicitly or implicitly, within the operational definitions of behaviour.

The parameters of frequency, duration, and IRT have been identified as the variables controlling the differential validity of data obtained by varying observational methods (Green et al., 1982; Milar & Hawkins, 1976). To anticipate discussion of Study 2, it was hypothesised that these same parameters would exert some control on the validity of data obtained from observation

samples of different duration.

Observational data collection

The method employed to measure behaviours can place constraints on the number of parameters which can be extracted from the behavioural record obtained (Green & Alverson, 1978; Sackett, 1978). In addition, the validity of data obtained by any particular recording procedure depends on the parameters of the behaviours being measured (e.g., Green & Alverson, 1978; Mansell, 1985). Instead of choosing a measurement strategy by consideration of validity most researchers seem bound by arbitrary convention, educated guesses, or convenience (Rojahn & Kanoy, 1985; Sanson-Fisher, Poole, & Dunn, 1980). There have been exceptions (e.g., Jones et al., 1975; Van Biervliet, 1982) but, in general, behaviour analysts rarely present any rationale for their choice of measurement method.

Interval recording, time sampling, and event recording have been the methods used most frequently (Kelly, 1977). Perusal of the literature indicates that this is still the case. Paper, pencil, and a time-keeping device are the only pieces of equipment required although the data recording can be assisted electronically (e.g., Van Biervliet, 1982), or by a considerable variety of ingenious means (for review, see, Bates & Hanson, 1983). With event recording a tally is kept of the number of times a behaviour occurs in the session. Both interval recording and time sampling involve division of an observation session into samples or intervals defined by their duration. Thus, an observation session is broken up into discrete samples. For

example, a 10-minute session could be divided into 10 one-minute, 60 10-second, or 200 3-second samples. Generally, the presence or absence of a category of behaviour within each sample (interval recording) or at the end of each interval (time sampling) is recorded, i.e., tick/cross or yes/no on a data collection sheet. Many variations on both sampling methods have been described and have been reviewed in behavioural texts (e.g., Kazdin, 1982; Sulzer-Azaroff & Mayer, 1977).

The differential validity of these methods has been shown to be controlled by the interaction between the units of measurement, i.e., the temporal parameters of the method, and the temporal parameters of the behaviour observed (for review, see, Rojahn & Kanoy, 1985). As an extreme example, if the observation procedure required that a behavioural event be marked as occurring only if it continued throughout an interval (the whole interval method) but the duration of the event was always less than the interval length, then the record would show that the event never occurred whereas there could actually have been many events. To avoid this and lesser threats, validity of an observation system can be determined empirically as demonstrated by Sanson-Fisher et al. (1980) or by consulting the literature on the subject (e.g., Rojahn & Kanoy, 1985). However, these methods do not allow for assessment of changes in validity which are inevitable when the parameters of behaviour change as is the aim in intervention studies.

Data obtained by these so-called pencil and paper systems are usually reported as modified frequency, the percentage of

samples in which the behaviour occurred. Unless the validity of the observation procedure was determined continuously the relationship between the data reported and the basic parameters of behaviour must be regarded as ambiguous (Sackett, 1978). In the case of event recording, the frequency of behaviour can be known but not the durations or IRTs. For interval recording and time sampling, none of these measures can be determined with known accuracy (Green & Alverson, 1978).

There are three methods to record behavioural events which can be employed to reduce substantially the problems of ambiguity in interpretation of data: video recording, electromechanical real-time recording, and real-time recording using human observers. All have advantages and disadvantages.

Video recording can produce a permanent record of behaviour for subsequent analysis and re-analysis (e.g., Linscheid et al., 1984; Powell, Martindale, & Kulp, 1975). In principle the basic parameters of behaviour can be measured from video recordings although difficulties may arise from the restricted field of view provided by fixed cameras, reactivity to hand-held cameras, and lighting problems which can make detection of subtle responses difficult (e.g., Mudford, 1985, Project 1).

Real-time recording of behaviour produces continuous measurement. This can be contrasted with the discrete sampling procedures discussed so far. With continuous measurement the recorder responds only when a subject's behaviour changes rather than at the completion of a designated interval. With real-time recording, only behaviour changes along with the real (clock)

time when these events occurred need be recorded. From data collected in this manner a representation of the stream of behavioural events is preserved to allow quantification of frequency, durations, and IRTs. Here lies the major advantage of real-time recording.

Electromechanical recording of behaviour involves the detection and permanent recording of behavioural responses directly, i.e., without a human observer. The general principle of automatic observation systems is that of using mechanical transduction to record the occurrence of behaviours. Microswitches, photocells, and ultrasonic detectors have been employed as transducers. Examples of applications of this measurement method as applied to detection and recording of motor responses have been reviewed by Pfadt and Tryon (1983).

The advantage of using electromechanical recording is that threats to validity due to observers can be eliminated. Where possible, automatic recording should be the method of choice. However, most responses of interest to applied behaviour analysts can not yet be reliably detected by these methods. No doubt future technical advances will increase the range of behaviours which can be so measured.

Observer-operated real-time recording systems

Given the present technical limitations on the use of fully automated recording to obtain real-time data, the best compromise is to employ human observers with the apparatus to record behaviour in such a form. The degree of compromise is, as yet, unclear as the validity of these recording systems has yet to be

assessed (Klesges, Woolfrey, & Vollmer, 1985; Schinke & Wong, 1977).

When observers record behavioural data in real-time an alphanumeric code is assigned to each of the behaviour categories into which the performance of the subject is divided. When the subject's behaviour changes to a different category a new code is entered into the apparatus which stores the code and the time for later retrieval and analysis. Since analysis by computer is most efficient, accurate, and time-economical, only those systems which are compatible with computers will be discussed. [It should be noted that less technically advanced methods have been used, e.g., speaking codes into a portable tape recorder (Landesman-Dwyer & Sackett, 1978); writing sequences of events and times (Bijou et al., 1968); push-button activated pen recorders (Harmatz, Mendelsohn, & Glassman, 1975; Lovaas, Freitag, Gold, & Kassorla, 1965); and real-time portable printers (Buckley, Frazer, & St. Amour, 1979)].

Computer compatible systems use a keyboard for the observer to enter codes for behaviour changes and a computer to analsye the data. Systems differ according to the storage medium which preserves the data between keyboard and computer. According to White (1971), the earliest storage media were papertape and punched cards. White (1971), Magyar and Fitzsimmons (1979), and Sackett, Stephenson, and Ruppenthal (1973) have described different systems which employed magnetic audiotape as the storage medium. Torgerson (1977) has developed a solid state collection and storage device, the Datamyte. Description and

review of several of these devices occupy half an issue of Behavior Research Methods and Instrumentation, vol. 9(2), 1977.

In comparison with pencil and paper recording these devices have the disadvantages of commercial unavailability or high cost and, potentially, are susceptible to mechanical failure and physical damage resulting in high maintenance and repair costs (Schinke & Wong, 1977). However there are potential savings in observer training time, observer fatigue, quantity of data which can be recorded, and validity of data. Empirical evaluation of these presumed benefits has yet to be undertaken.

The real-time collection and storage systems reviewed can be described as dedicated to that function alone. The expense involved in their acquisition and maintenance has not been considered to be justified by some researchers (e.g., Van Biervliet, 1982) and accusations of technical pretension have been levelled at their use by others (Rojahn & Kanoy, 1985). It can be argued, however, that these types of dedicated apparatus are no longer required. If portability is not a concern, then desk-top personal computers can be programmed to collect, store, and analyse real-time data (Flowers & Leger, 1982). Most researchers and many clinicians either own or have ready access to such equipment. Some suitable programmes are available at little or no cost (e.g., for TRS-80 computers: Balsam, Fifer, Sacks, & Silver, 1981; Deni, Szijarto, Eisler, & Fantauzzo, 1983; Koontz, 1982. For Apple II computers: Flowers, 1982; Moss, 1984). If portability is required, recently available hand-held computers could be used for the same purpose (e.g., Psion

Organiser). In either case the expense is not dedicated to data collection as the computers can be used for a wide variety of other tasks.

Rationale for this study

Empirical studies of the validity of the type of information obtained by direct observation have been conducted with data acquired from a variety of sources other than from direct observation of potential or actual subjects. For example, computer-simulated pseudo-behaviour has been generated by some (e.g., Harrop & Daniels, 1986; Rojahn & Kanoy, 1985) and videotapes of manufactured behaviour have been used by others (e.g., Green et al., 1982; Powell, Martindale, Kulp, Martindale, & Bauman, 1977). Behaviours which have been chosen for convenience of recording rather than those of intrinsic interest to behaviour analsyts have also been measured (e.g., Powell, Martindale, & Kulp, 1975).

Distinct advantages are available to researchers using these artificial methods. The important parameters of behaviour are specified by those employing computer-simulation or manufactured data and are easily quantified and briefly reported when convenience is the criterion for choice, e.g., the electromechanical recording of the in-seat behaviour of a secretary (Powell et al., 1975). An advantage to the science of behaviour analysis of employing such artificial databases from which to judge the validity of various observational procedures is that the effect of variation of each parameter can be demonstrated quite simply. It should be pointed out that

theoretical studies (e.g., Ary, 1984; Milar & Hawkins, 1976) can serve the same purpose although the mathematical derivation of relations between the parameters of behaviour and validity may be poorly understood or ignored by the majority of behaviour analysts.

The real-time data collected for this thesis was directly input to an IBM personal computer by observers and the raw data were stored on floppy disk. Analysis was performed on the same apparatus thus avoiding the need to transfer data from one storage medium to another as is required when using dedicated data collecters.

The present study has been separated from Study 2 for two reasons. The first is, for the subsequent experiments, Study 1 can be seen as an extension of the Method section in that it describes the database more fully than the description of subjects and setting. Secondly, while quantified descriptions of the activities of students in some educational settings have appeared (e.g., Green et al., 1986; Ysseldyke, Thurlow, Mecklenburg, and Graden, 1984), the performance of physically disabled profoundly mentally retarded adults in training has not been assessed.

Method

Participants

The subjects were selected from residents of Nikau Villa, Templeton Hospital and Training School, Christchurch. This ward catered for 27 long-term admissions of adult physically handicapped mentally retarded people. Five of 10 residents who

attended the ward's training area were selected for observation because their attendence was most consistent. These five did not regularly leave the group for physiotherapy, occupational therapy, or speech therapy. Table 1 presents basic information on the subjects. Reported retardation levels were obtained from

---

Insert Table 1 about here

---

clinical assessments conducted annually and attached to each resident's ward file. All could eat independently with a spoon. None had intelligible speech or manual communication skills but some staff claimed to understand some of one subject's (AM) vocalisations for "Yes" and "No". None had independent toiletting skills although two (GT, AM) kept their clothes dry through routine toiletting initiated by, and with help from, staff. Permission to observe these residents had been received from the institutional authorities and to videotape samples of the behaviours of MC and GT from their parents.

Staff members who worked with the training group were three training officers, specialist teachers of severely and profoundly retarded people (Ahrens, 1986). One was a trainee with seven years' experience working with mentally retarded people, one certificated with five years' experience, and one assistant training officer who had one year's experience. All were in their 20s. At times other staff were also present, e.g., psychopaedic nurses and assistants, physiotherapists and others who were helping, visiting residents or staff, or passing through the

Table 1

Basic information about subjects

	Subjects				
	MC	MD	PD	GT	AM
Age (years)	30	35	34	26	34
Level of retardation (Grossman, 1983)	all	profoundly		mentally	retarded
Years in institution	26	28	28	6	21
Etiology	meningitis (16 mths.)	congenital	congenital	?birth injury	birth injury
Medication (dose in mg./day)	epilim (1600) tegretol (600) phenobarb (30) (for epilepsy)	---	----	stelazine (6) (for maladaptive behaviours)	cogentin (2)

Table 1/cont.

Weight (kg.)	48	n/a	n/a	44	53
Additional handicaps	hemiplegia	diplegia	diplegia	moderate spastic quadriplegia	diplegia
Mobility	can walk but is usually assisted by staff and can shuffle on floor	propels wheel-chair and rolls on floor	propels wheel-chair and crawls	walks with support of staff or furniture	propels wheel-chair

room.

Setting

Observations were made in the training area of Nikau Villa. The main room used for training was spacious (14 m x 7 m), warm (from underfloor heating), and carefully decorated, e.g., silhouettes of residents, their crayonned scribblings, some indoor plants, and various colourful hanging mobiles. Generally the north-west end of the room was used during morning sessions (8.30 - 11.00 a.m.) for which the residents were seated on chairs or in wheelchairs at their own fold-up tables along the walls. Attached to the wall above each table was a written profile of a resident detailing abilities, likes and dislikes, idiosyncratic communication methods, and tried-and-true methods of handling behaviour problems.

Equipment in the training area included a staff-operated radio and stereo record player and for the residents: paper and crayons, peg boards and form boards, simple jigsaws, soft toys in abundance, toy cars, Lego, balls and bean bags (for throwing). During the mornings this equipment was supplied to residents by staff although residents did have sufficient mobility to approach peers or equipment and all subjects were noticed, on occasion, to do so. Training staff circulated around the group assisting its members in various ways, e.g., helping them brush their hair, giving manual guidance with the equipment provided, and talking to residents in addition to 'caring' activities such as changing clothes, lifting and moving residents. At about 9.45 a.m. residents were helped with pouring, or served, a cup of cordial

and given a biscuit after which they were assisted to the toilet or had their clothes changed, as neccesary. Next, staff had a tea break in the adjoining room while maintaining cursory surveillance of residents. Activities were the same as before morning tea for the last half-hour of the session.

Afternoon sessions (1.00 - 3.30 p.m.) were usually conducted with residents surrounding the large table at the south-east end of the room. Activities included finger painting, sticking papier mache to inflated balloons, water play, and cooking (i.e., residents stirred ingredients for pikelets and the like).

At about 2.45 p.m. the residents were given drinks and were toiletted; staff then had a tea break. Most of the last 20 minutes of the afternoon session was taken up with tidying up the room by staff.

Equipment

Real-time recording of behaviour was acheived by using an IBM PC portable computer programmed in BASIC. The computer had a memory capacity of 512 kilobytes and two disk drives. Although described by the manufacturers as portable, the computer weighed 13.5 kgs and, in use, had dimensions of 65 cm x 50 cm x 20 cm. It required a mains power supply for operation. Software written to store and analyse the data will be briefly described where neccesary in the appropriate Procedure sections. Appendix 1 contains a printed copy of the programmes used.

Observation categories and codes

An exhaustive and mutually exclusive set of categories was used to describe subjects' behaviours (Sackett, 1979). The means

for selecting behaviours for observation were similar to those described by Harmatz et al.(1975). Notes were taken during seven hours of informal observation. The notes were then scrutinized for classes of behaviour which appeared to have face validity in providing an assessment of the subjects' activities in the training area. Behaviours of possible interest which were entirely dependent on the behaviour of staff were combined into broader categories. For example, subjects in this setting could eat and drink only when a staff member gave a subject some food or a cup of drink; having food or drink in the hands was included in the 'hands on objects' category of behaviour. Categories were chosen which could occur at any time during the observation session, i.e., neither peculiar to a particular part of the session nor dependent on staff assistance.

Each category was arbitrarily assigned a single digit-code and the experimenter conducted two 150-minute sessions with the equipment to check for potential difficulties with categories and coding. Following refinement, six categories were selected (see Table 2).

Insert Table 2 about here

To obtain mutual exclusivity of categories a priority of coding was devised (Sackett, 1978; Sanson-Fisher et al., 1979). For example, if the subject was receiving manual guidance from a staff member to manipulate an object, the code for social staff was entered in priority over hands on objects. The coding

Table 2

Behavioural categories, definitions, and codes

Code	Category	Definition
1	social staff	Being attended to from a distance of 1 metre or less; touching or being touched by staff; hands on same object as staff; vocalizing at or being vocalized at by staff.
2	social peer	Touching or being touched by peer; hands on same object as peer; vocalizing at or being vocalized at by peer.
3	self-move	Moving own wheelchair, crawling, walking, getting into or out of chair.
4	inappropriate	Engaging in stereotypy, self-hitting or scratching, inappropriate undressing, banging furniture or objects, damaging objects, yelling.
5	hands on objects	Not merely resting hand(s) on objects but doing something with object, i.e., moving object or moving hand(s) over object. Object - doesn't include furniture
6	passive	e.g., asleep, gazing, staring, looking around.

Note: Priority of coding : 2, 1, 5, 3, 4, 6

priority system used is noted in Table 2. It can be noted here that the priority coding was never required for social peer over social staff, i.e., these categories never occured simultaneously (although there was no reason why they could not). Similarly, inappropriate never occurred in conjunction with another category and hands on objects was never coded in priority to self-move. However social staff overruled self-move, hands on objects, and passive not infrequently. Occasionally social peer was coded when this category occurred at the same time as hands on objects. The main effect of the priority system was to make the categories self-move, hands on objects, and passive independent of staff assistance or attention.

Because the coding system was developed over 12 hours of informal and semi-formal observation and formal data collection extended over 25 hours, it was thought possible (even likely) that some uncodeable event would occur which had not been forseen. For example, an epileptic seizure could hardly be coded as inappropriate or passive and a physical attack on a peer or staff member could not sensibly be coded as social. In such an event, which never occurred anyway, observers were instructed to press the keyboard's "T" which displayed the time on the visual display unit (VDU or screen) and manually record the type of event and its start and finish time.

Observers and their training

Ten undergraduate students enrolled in a third year course in Applied Behaviour Analysis and the experimenter, a graduate student in psychology, acted as observers. The observers were

informed of the hospital's rules regarding confidentiality of information about residents and were read the section on observing human subjects from the ethical guidelines approved by the New Zealand Psychological Society (1986).

The observers' first exposure to the real-time recording system and the codes employed was through a one-hour laboratory session in which observers coded 10-minute videotaped samples of two of the subjects' behaviours. Pairs of undergraduate observers were allocated to in vivo training and observation sessions.

In vivo training was conducted for pairs of observers after they had become accustomed to the setting and had been introduced to potential subjects. Observers then selected a subject who had not been chosen by other observer pairs and were trained to observe that subject. While observing the subject, the experimenter discussed the coding system as applied to that subject, e.g., the topography of stereotypy (if any) in that subject's behavioural repertoire. Next, observers were seated at the computer keyboard and entered codes for behaviour categories as the experimenter spoke the category name. The experimenter gave feedback on accuracy of coding. Reliability was assessed after 10 minutes of coding and was provided to observers as further feedback.

Observers next completed a 10-minute session in which verbal communication about appropriate codes was encouraged. Again, feedback on accuracy was provided during the session and on reliability immediately after its completion. A further 10-minute sample was coded by both observers without communication between

them. Feedback on accuracy and reliability was provided as in the previous session.

Finally, one or two further 10-minute sessions were conducted without communication or feedback during the session. The criterion for completion of training was a 10-minute session in which reliability was acceptable (defined as kappa > .75) and the experimenter judged that the recording was accurate from simultaneous observation of the subject's behaviour and the observers' coding responses.

All observers were trained to criterion in less than one hour of in vivo training. Apart from the experimenter all observers were naive to the purpose of the observations.

Procedure

Each subject was observed for an entire training day, i.e., the five hours spent in the training area. Observations started at a minute or so before 8.30 a.m. and 1.00 p.m. and extented for two and a half hours (150 minutes) in each session. Sessions are described as being of 150 minutes duration although all exceeded this by between 36 and 148 seconds.

The ward's charge nurse and the training officer in charge of the training area were informed who was to be observed on which day. This was neccesary to ensure that the subjects were ready in the training area at the session starting times and remained in the setting for 150 minutes. Although staff and subject reactivity to observation was not formally assessed, it was noted that no participant behaved differently during observation. Staff were not aware of the behaviour categories

being recorded nor of the purpose of the study. They were informed, correctly, that the behaviour of individual staff members was not being assessed.

A few minutes before the start of observation sessions the computer was set up and the data recording programme INPUT was loaded. The output file, on floppy disk, was given a unique name, the subject's initials plus "1" for a.m., or "2" for p.m. sessions. At the session start-time the return key was pressed and the primary observer, either the experimenter or one of the observers trained with that day's subject, entered the code for the category of behaviour in which the subject was engaged. The primary observer entered codes via the numeric keypad on the right of the keyboard. Observers had a printed list of codes and definitions available during sessions.

When the subject's behaviour changed to that defined by a different category the observer entered a new digit-code. Because categories were mutually exclusive, codes were input only to signal the start of a category's occurrence. Codes input were displayed on the VDU to enable the observer to check which code was on. When possible primary observers alternated (i.e., were interchangeable) about half-hourly although observer fatigue was not found to be a problem even for a continuous 150-minute session.

If, during observations, the subject was not visible from where the primary observer was sitting another observer would follow the subject and use hand signals to indicate a change of code. This occurred when the subject went to the toilet, or a

clear view of the subject was blocked by other people, or the subject was facing away from the observers. When no second observer was available and the subject was not visible from the computer keyboard, the primary observer recruited one of the training officers to input codes signalled by the observer. This did not provide an opportunity for the staff to learn the behavioural categories being recorded as only the observer could see the subject on these occasions.

There was no code to indicate an error in recording as has been used in some electronically assisted recording systems (e.g., Sanson-Fisher et al., 1979; Van Biervliet, 1982). It was considered that such a code would cause more difficulties than it would overcome. For instance, if a sequence of codes was "6" at time #0, "4" at time #1, "error code" at time #2, "5" at time #3 would the software be instructed to place code 6 at time #1 and time #2 or code 5? Also, by the time the observer had entered the error code then the correct code the subject's behaviour may have changed again in which case employment of an error code would lead to even greater inaccuracy of coding. Simpson (1979) came to much the same conclusion in suggesting that errors would be reported with disagreements in reliability analyses. In the present study, observers were instructed that if they made an error they should enter the correct code at that time as soon as the error was noticed. In fact, errors were rarely noted probably due to the simplicity of a six-code recording system and the disabling, by software, of all keyboard keys which did not correspond to an allowable code.

The end of the session was signalled to the computer by an exclamation mark (!). When "!" was input the entries which had been made in the previous 150 minutes were stored on diskette along with the duration of each entry; that is, the number of seconds between entries.

Programming note. Codes were input to an INKEY loop then the ASCII bit-code was checked to ensure that the code was allowable and was not the same as the previous code input [to avoid multiple entries of redundant data should, for example, an observer hold a key down (White, 1971)]. Time of input was obtained from the computer's internal clock by the "variable = TIME$" function. Each code and its corresponding TIME$ was stored in memory until the end of the session. The start-time and the number of seconds between code inputs were stored rather than the TIME$s to economise on diskette storage space, i.e., two-byte integers rather than eight-byte strings or 'seconds since midnight' as four-byte single precision numbers.

Reliability

The reliability of observations was assessed by using two observers simultaneously and comparing their coding behaviour using the statistic kappa (Cohen, 1960; Hollenbeck, 1978). The second observer sat to the left of the primary observer and entered codes via the keyboard keys Z, X, C, A, S, D which corresponded to the first observer's codes 1, 2, ... 6, respectively. Neither was informed of the other's codes although the experimenter who acted at times as the primary observer and sometimes as the secondary observer knew both sets of codes.

Observers were instructed not to converse about coding during the observations. Non-compliance was not observed.

Reliability observations were spaced throughout a session and occupied between 26% and 34% of the whole training day for each subject. Sometimes it was impossible to schedule a second observer for a session (150 minutes). In these cases the experimenter made observations for the entire session and reliability checks were not conducted. If a reliability check was in progress when the subject disappeared from view the second observer immediately terminated the check, by inputting "&".

A value of kappa at .75 was selected as the minimally acceptable level of reliability, following the recommendation of Gelfand and Hartmann (1984). Kappa was computed separately for each reliability check; the second-by-second comparison procedure and algorithm detailed by Hollenbeck (1978) having been programmed into the computer. There was no feedback to observers on reliability during observation sessions.

Obtained values of kappa are presented in Table 3 along with

Insert Table 3 about here

---

the duration of reliability checks. In 29 of the 30 reliability checks kappa exceeded .75.

Data analysis

Each of the 10 whole session records (150 minutes) was analysed by two computer programmes. Relative duration of each

Table 3

Obtained values of kappa and percentage of session during which reliability was checked

	Subjects				
	GT	AM	MD	PD	MC
Morning sessions (kappa)	.94 .78 .79 .89	.96 .76 .89 .89	no second observer available	.83 .93 .92 .33 .91	no second observer available
% of morning	38%	41%	0%	53%	0%
Afternoon sessions (kappa)	.79 .82 .84 .85	.85 .90 .86	.96 .84 .84 .94 .87	no second observer available	.92 .97 .92 .95 .93
% afternoon	31%	14%	66%	0%	62%
% of all observation	34%	28%	33%	26%	31%

of the six codes was computed as a byproduct of programme EX1B2 (see Appendix 1) used in Study 2. This measure was of the percentage of the whole session during which the code was on. Codes which had a relative duration of less than .7 percent of a particular session were not included in any further analysis. It was judged unlikely that anyone would measure such low relative duration behaviours by real-time or any time sampling technique, since event recording would be the likely method of choice (Sulzer-Azaroff & Mayer, 1977).

The frequency and distribution of individual event durations (i.e., how long codes were on) and IRTs (i.e., how long between events of the same type) were obtained from the programme IRTS (see Appendix 1). Frequency was defined as the number of entries of a particular code in a session, i.e., the total of occasions when the first observer judged that the subject's performance changed to a category of behaviour represented by the code.

Expected average IRT was derived for each code in each session from the relative duration and frequency data. This was manually calculated by multiplying 100 minus the relative duration by the session length (in seconds), then dividing by 100 times the frequency. This measure was produced for comparison with the obtained average IRT. If the expected average IRT exceeded the obtained average IRT then there is an indication that all instances of a particular code in a session were temporally clustered.

Average durations were manually calculated by multiplying relative duration by the session length (in seconds), then

dividing by 100 times the frequency.

The programme IRTS printed individual code durations and IRTs in the form of distributions across 10-second blocks to reduce the volume of data which could have been produced for these measures. That is, the number of durations of between 1 to 10 seconds was counted, the number between 11 to 20 seconds, and so on. IRT distributions were similarly counted. Obtained average IRT was calculated by taking the mean of the individual IRTs for each code in every session. It should be noted that, as the mean was calculated from blocked data, the obtained average IRT is an approximation: Each mean was calculated from a number of datapoints equal to the frequency and each datapoint was approximate to within plus or minus five seconds.

Obtained and expected average IRTs were compared by calculating the difference between expected and obtained IRT as a proportion of the expected IRT, i.e., (mean expected IRT minus mean obtained IRT) divided by mean expected IRT.

Results

Relative durations of codes across sessions are presented in Table 4. Although variability across sessions and subjects can be noted some general observations can be made. For social

---

Insert Table 4 about here

---

interaction with staff (code 1) relative duration varied between 9% and 31% of a session, the mean being 14% which, in terms of

Table 4

Relative durations of codes (percentage of session) in morning (am) and afternoon (pm) sessions

		Subjects				
		GT	AM	MD	PD	MC
Code 1	am	13.0	30.8	12.4	10.3	8.8
	pm	11.3	13.3	16.8	15.4	27.1
Code 2	am	6.1	>0	11.1	1.4	1.4
	pm	2.2	>0	0.7	6.1	1.0
Code 3	am	5.9	11.8	4.8	1.5	0.2
	pm	13.8	4.5	11.3	4.9	0.1
Code 4	am	0.0	0.0	1.1	0.6	1.1
	pm	0.0	0.0	0.2	0.4	0.2
Code 5	am	36.4	46.2	26.8	56.6	2.9
	pm	14.6	41.4	35.5	25.5	1.7
Code 6	am	38.6	11.1	43.9	29.6	85.5
	pm	58.1	40.8	35.6	47.6	70.0

absolute duration, was 21 minutes of a 2.5 hour session. Inappropriate behaviour (code 4) occurred at low levels (1% or less) and not at all for GT and AM. Overall the two predominant behaviour categories were hands on objects (code 5) with a mean (excluding MC) of 36% and passive (code 6) with a mean of 46%. Thus, on average, subjects spent 69 minutes in each session doing nothing. If the time when being attended to by staff is removed, that is, independent behaviour is examined, the mean for passive increases to 54% with a range from MC at 95% to 33% for AM for their whole training days.

Data for frequency, event durations, and IRTs are presented in Table 5. Each row of data is preceeded by the initials of the subject followed by 1 for a.m., 2 for p.m. sessions and the behaviour category code. The data for codes 2-6 (subjects'

Insert Table 5 about here

---

behaviour independent of staff) have been grouped according to relative duration for ease of reference from Study 2. The explanation for the generally wide discrepencies between average duration of events and maximum durations is that the duration data were skewed towards short durations with some in the 1 to 10 second block for nearly every code in every session (there was one exception). A similar explanation can be given for the difference between average and maximum IRTs.

The comparison between expected and obtained average IRTs in Table 5 requires elucidation. Some small negative values which

Table 5

Quantified description of the database sampled in Study 2 grouped according to relative duration of codes in sessions: frequency, average duration (seconds), maximum duration (seconds), expected and obtained average IRTs (seconds), maximum IRTs (seconds), and expected versus obtained average IRT comparisons.

Relative duration	Subject session code	Frequency	Durations		IRTs			
			mean	max.	exp.	obt.	max.	comparison
>50	MC1-6	50	155	1825	26	28	285	-.08
	MC2-6	59	107	845	46	47	895	-.02
	GT2-6	93	56	595	41	40	395	.02
	PD1-5	43	119	1040	91	80	685	.12
Means		62	109	1076	51	49	565	.04
25-49.9	GT1-5	93	35	335	62	63	645	-.02
	GT1-6	91	39	305	61	61	435	.00
	MD1-5	57	43	505	102	103	615	.01
	MD2-6	82	40	295	71	71	625	.00
	AM1-5	32	129	2630	154	137	1155	.11
	AM2-5	54	68	585	100	99	1115	.01
	AM2-6	56	66	465	96	94	585	.02
	PD2-5	62	38	325	109	110	745	-.01
	PD2-6	121	36	365	39	39	475	.00
	PD1-6	65	42	445	98	99	1075	-.01
	MD1-6	109	36	345	54	47	605	.12
Means		75	52	600	86	84	734	.02

Table 5/cont.

10-24.9	GT2-3	36	35	115	217	224	2285	-.03
	GT2-5	42	32	325	184	181	1825	.02
	MD1-2	29	34	225	280	249	3815	.11
	MD2-3	26	38	105	312	161	2895	.48
	AM1-3	37	29	135	217	219	4355	-.01
	AM1-6	44	23	155	184	186	2735	-.01
Means		36	32	177	232	203	2985	.13
3-9.9	GT1-2	37	15	55	230	180	1985	.22
	GT1-3	43	13	45	198	111	735	.44
	MD1-3	19	24	105	456	483	4095	-.06
	AM2-3	23	20	75	377	387	3105	-.03
	PD2-2	20	27	235	430	393	2855	.09
Means		28	20	103	338	311	2555	.08
.7-2.9	GT2-2	19	10	45	468	470	3415	.00
	MD1-4	9	11	55	1004	980	4785	.02
	PD1-2	17	7	15	531	333	3205	.37
	PD1-3	8	17	75	1118	1029	3965	.08
	MC1-2	9	14	55	1005	27	95	.97
	MC1-4	16	6	35	565	513	3225	.09
	MC1-5	5	52	175	1772	30	55	.98
	MC2-2	4	23	45	2237	472	1345	.79
	MC2-5	9	17	65	985	723	3505	.27
Means		11	17	63	1076	509	2213	.53

Table 5/ cont.

code 1	MC1-1	21	39	285	396	372	1825	.06
	MC2-1	44	55	895	150	132	1890	.12
	MD1-1	37	29	155	217	216	2335	.00
	MD2-1	65	24	275	116	116	1095	.00
	AM1-1	48	58	695	131	131	2675	.00
	AM2-1	42	28	125	189	191	1725	-.01
	PD1-1	33	27	105	249	238	2375	.04
	PD2-1	87	16	115	89	89	1415	.00
	GT1-1	46	25	175	171	174	2485	-.02
	GT2-1	55	18	195	147	142	1855	.03
Means		48	32	302	186	180	1968	.03

appear will seem impossible to the computationally alert reader. The average IRT data were, as previously explained, approximate. These small errors will not affect further discussion. In some comparisons the figure is positive and non-negligable, e.g., session MD2 code 3 at .48; session MC1 code 2 at .97, and code 5 at .98. These figures indicate that all occurrences of these codes appeared in a relatively small part of the session, e.g., in MC1, code 2 occurred for 1% of the session but all nine instances were within a four-minute portion (subtracting 1 from the frequency then multiplying by the average obtained IRT).

To attempt to summarise the data in Table 5: Subjects behaviour was highly variable. Often behaviour categories were re-entered relatively frequently (i.e., short average IRTs) for relatively short time periods (i.e., relatively short average durations). Sometimes, however, events lasted for relatively long periods (i.e., long maximum durations) and re-entries were widely spaced (i.e., long maximum IRTs). Occasionally for low relative duration codes all instances of the code were clustered (i.e., large positive expected versus obtained average IRT comparison values).

Discussion

Results presented in this study show behavioural similarities and differences across subjects who were continuously observed during morning and afternoon training sessions. Staff interaction, averaging 14% of a session, may be judged as occurring at a desirably high level although the definition for the social staff category has no qualititive

implications. For subjects' behaviours independent of staff interaction, the passive and hands on objects categories predominated although one subject, MC, exhibited very low levels of the latter. Inappropriate behaviours occurred for 1% or less of sessions for all subjects. Social interaction with peers was observed at low levels, 2% or less, in 7 of the 10 sessions but at higher levels, 6 - 11%, in one session each for GT, MD, and PD. Self-movement varied across subjects and sessions from a maximum of 14% in GT's afternoon session to less than 1% of MC's morning and afternoon.

Social validity of behavioural definitions

A serious problem with this study concerns lack of social validity of some of the definitions of behaviour categories (Kazdin, 1982; Wolf, 1978). Traditionally, this type of validity was subsumed by the heading of content validity (Cronbach, 1970). It has also been described as ecological validity (Foster & Cone, 1980). The question is, simply, how much relevence do the behavioural definitions (and the values implied by them) have to those concerned or affected by a study? In this case those directly concerned were the subjects, their staff, and their parents. Indirectly concerned or affected are the consumers of this report. As stated, the social staff category was defined without regard to quality or type of interaction. For example, if the subject being observed reached out and touched a member of staff the code for social staff was entered regardless of whether the staff member appeared to even notice the attempted initiation of interaction. This same problem was apparant for the social

peer category and for the hands on objects category. In the latter category merely fiddling with an object, e.g., a piece of Lego, was entered as hands on objects. So was turning the pages of an upside down magazine, rearranging soft toys, and eating a biscuit.

For behavioural definitions to describe interaction with staff the eight categories defined by Repp and Barton (1980) or, at least, the four defined by Van Biervliet (1982) would be more meaningful than the single category used in the present study. For socially valid assessment of manual performance (hands on objects in this study) at least two categories would need to be defined, i.e., functional and non-functional with respect to the aims of training as used by Green et al. (1986). Although the relatively high proportion of hands on objects for four of the five subjects may seem to indicate desirable activity, very few of the events so coded could have been described as functional given the equipment provided to the subjects.

Inter-observer reliability

Some further problems with this study as an attempt to contribute to the database from which decisions about the needs of and services for physically handicapped profoundly mentally retarded people concern the issue of interobserver reliability. When checks for reliability were being conducted both observers were entering codes through the same keyboard. They were sitting next to one another and each could see, by looking at the VDU, which codes were being entered. Although they were not informed of the other's codes, it would not have been difficult to work

out what category was represented by what code with only six behaviour categories in the observation system. This is not to imply that this actually happened but the method described allows for this possibility.

From observation of the coding of two observers simultaneously one aspect was striking. Although observers were oriented towards the subject, rather than each other or the VDU, the click of an entry-key by one observer seemed to alert the second observer to decide whether the subject's behaviour had indeed changed to a new category. It seemed that one observer in each pair assumed dominance on decision-making to enter a new code but not which particular code. The effect on estimates of reliability of observers cuing one another can not be ascertained. It could have increased, decreased, or not affected reliability. Without doubt the believability of the data would have been enhanced by having observers coding on spatially separate machines or keyboards but limitations of space and resources precluded such a procedure in this study. However, with only six categories of behaviour to code and no category requiring inference or qualititive judgement, accuracy was likely to be enhanced (Kazdin, 1977) and was judged to be high from observation of the coders, although this was not measured.

Kazdin (1977) has recommended that reliability checks be conducted unobtrusively as reliability has been found to improve when observers were aware that their behaviour was being assessed. Other advice given to obtain accurate estimates of reliability has included the introduction of newly trained

observers or videotaping of sessions to guard against observer drift and computation of reliability estimates by an independent party to avoid inflated measures (Kazdin, 1977). On the positive side, the calculation of reliability estimates was performed by computer, an independent assessor, provided its programme was not biased to inflate estimates. Against the method used, observers were not totally independent during data collection (Harris & Lahey, 1982); observer drift was not assessed although all observers except the experimenter could be considered as newly trained; and reliability checks were far from unobtrusive.

The last aspect to be raised about the issue of interobserver reliability concerns the use of the coefficient kappa as the measure of reliability (Cohen, 1960; Hollenbeck, 1978). There are two types of kappa, weighted and unweighted. The weighted version has been described by Cohen (1968). Apparantly weighted kappa has been used to assess the reliability of observational data obtained by an interval sampling method (Alevizos et al., 1978) and from real-time records of behaviour (Poole et al., 1981). It is not clear from reading these examples how weights have been applied as the only indication of their use is the reference to Cohen (1968). It seems instead highly likely that unweighted kappa (Cohen, 1960) has been used as described by Hartmann (1977) and House et al. (1981) and that the apparant examples result from inaccurate referencing.

In the present study unweighted kappa (kappa) was used as a measure of reliability. The formula for calculation of kappa is:

$$\text{Kappa} = (Po - Pc)/(1 - Pc) \quad \ldots \text{Equation 1}$$

where Po is the proportion of n units in which observers agreed and Pc is the proportion of n units for which agreement is expected by chance (n is discussed later in this section). This measure has been described as most suitable for real-time data (Hollenbeck, 1978). The type of data obtained, from independent subjects and with independent, exhaustive, and exclusive coding, complied with the assumptions required by kappa (Brennan & Prediger, 1981; Cohen, 1960).(The independence of observers, which is an assumption of all methods of estimating reliability, has already been discussed.) Also the agreement matrix produced as part of the computational procedure was anticipated to be useful for observer training. Specific sources of reduced reliability, e.g., confusions between categories, can be pinpointed (see Appendix 2). This was indeed useful for this purpose. Further, the use of kappa has not been proscribed or criticized by recent reviewers of reliability measures (House et al., 1981; Maclean et al., 1985). However it is unclear whether Hollenbeck's (1978) method which combines all behaviour categories to produce a single summary coefficient of reliability has the same acceptable status as Cohen's (1960) method which provides for separate kappas for each category. In a review acquired subsequent to data collection and analysis, Hartmann (1982) has recommended that, although calculation of a summary kappa is indeed possible, kappa should be calculated separately for each category: Hollenbeck's (1978) method was not included in any of the reviews cited.

The question of acceptable values for kappa is perplexing.

Sanson-Fisher and his associates (Poole et al., 1981; Sanson-Fisher et al., 1980) appear to have employed Cohen's (1960) method and tested a z-statistic derived from kappa for interobserver agreement being significantly greater than would be expected by chance. However, " . . . to know merely that kappa is beyond chance is trivial since one usually expects much more than this in the way of reliability in psychological measurement" (Cohen, 1960, p. 44). Hollenbeck (1978) offers no criterion for judging whether obtained kappa indicates adequate reliability when calculations are performed by his method. The criterion adopted in the present study, that kappa should exceed .75, was the most stringent recommended by Gelfand and Hartmann (1984). From the description of values of kappa in Landis and Koch (1977) the interobserver agreement data presented in this study (Table 3) can be described as fair in one case (k = .33), substantial in four cases (.76 - .79), and almost perfect in the remaining 25 cases (.82 - .97). It is not clear whether these categories apply to Hollenbeck's (1978) method.

To conclude discussion of reliability with real-time data especially with at least some long event durations, a feature of the present data, I suspect that kappa, calculated according to Hollenbeck (1978) may be an exceedingly lenient measure. Indeed Hollenbeck acknowledges that there is a potential difficulty in choosing n (the number of observations) for the computation (see Equation 1). Should n be the number of seconds in the reliability check, i.e., reliability assessed for second-by-second agreements (as used by Hollenbeck and in the present study), or the number

of behaviour changes in the check, i.e., input-by-input agreement? In other words, should the unit be a second or an event? Clearly, further research into appropriate statistics for determining estimates of reliability with real-time data needs to be undertaken.

Conclusions

Taking into account the lack of social validity of the behavioural definitions and the possibility of non-reliable data due to both the method for conducting reliability checks and the measure of reliability employed, it might be concluded that the data presented contribute little to our knowledge about mentally retarded people. While this may be so, it can be argued that, for the purposes of assessing sampling procedures in Study 2, the database was adequate. Regardless of the validity of the codes for assessment of performance the continuous records were of the behaviours of human subjects or, even if reliability is doubted to the extreme point, a perfectly accurate record of human observers' key presses. Table 4 and Table 5 quantify that performance in, admittedly, a more complex manner than parameters specified to computer simulation programmes. It is suggested, however, that the data presented are closer to the experience of behaviour analysts than that acquired from computer-simulation or manufactured behaviour.

Study 2

Validity of behavioural data obtained from observational sessions of different durations

A major universe in the study of the generalizability of behavioural assessment data is that of Times (Coates & Thoresen, 1978; Cone, 1977; Jones, 1977). Facets of the Times universe can be categorized according to temporal dimension expressing a variety of issues of concern to behaviour analysts. Assuming that exhaustive sampling through all the occasions when the behaviour of interest can occur is impossible, various questions arise from consideration of Times. For example: How representative is data from a session on one day compared with data obtained from a week of daily observational sessions? How valid is it to generalize from a week's data to a month? And so on.

Generalizability of data across time has usually been conceptualized in terms of the stability of data across observational sessions. Examples include recommendations that pre-intervention (baseline) observation sessions be conducted until the data looks stable (e.g., Parsonson & Baer, 1978) or has been statistically assessed as stable (e.g., Jones et al., 1975). Studies of the maintenance of behaviour change indicate that generalizability across long time periods (from weeks to years) can not be assumed (Stokes & Baer, 1977). Again, empirical assessment of maintenance has been judged by stability of the data across observation sessions.

Although data may indicate stability across days or months, the universe of Times has not been fully addressed. The representativeness of observational sessions with respect to the whole time of interest has been a neglected facet (Butcher, 1983). If this is unknown we can not estimate the validity of a single data-point, let alone a series, stable or not. The issue resolves to this: What duration should be chosen for observation sessions that allows reasonable generalizability to the whole time in which target behaviours can occur?

One answer to this question has been given which involves comparing data from the first and second halves of a session to check for stability by correlation (Hartmann, 1982) or by ANOVA (Mitchell, 1979). However, stability is not a substitute for representativeness.

The problem of selecting a duration for observation sessions has been widely recognised by writers on assessment methodology but suggested solutions have generally been vague or neither logically nor empirically justified. Haynes (1978) suggested consultation of previously published studies or derivation of session parameters statistically. The first suggestion would perpetuate the problem (if there is one) and the second could be performed only if the range of parameters of the behaviour under study could be known a priori. Random selection of times for observation has been suggested by Kazdin (1980) and one hour or several shorter periods by Kazdin (1984b). Bijou, Peterson, Harris, Allen, and Johnston (1969) also recommended a standard observation period of one hour for classroom observations.

Hartmann (1984) recommended minimizing costs and maximizing representativeness but does not suggest how this could be performed. Most writers have merely acknowledged the problem (e.g., Altmann, 1974; Goldfried, 1983; Wildman & Erickson, 1977) or suggested that session duration should depend on the purpose of the study, the nature of the data, and practical considerations (Bijou et al., 1968; Hartmann, 1984).

Among the texts surveyed, only Johnston and Pennypacker (1980) have suggested that all responses may have to be observed, i.e., the whole time of interest, at least temporarily, to assess empirically the generalizability of samples smaller than the whole. Again, it should be pointed out that this type of assessment need only be considered when it is practically impossible to observe through the whole time of interest. Of course, if the target behaviours can only occur for limited durations in specific settings, e.g., mealtimes, there is no problem. The whole universe of generalizability can practically be observed.

Some researchers have examined the validity of observation samples with respect to longer time periods. When assessing the behaviours of psychiatric in-patients, Alevizos et al.(1978) evaluated the representativeness of data obtained from two 15-second observations per day against a criterion measure obtained from 15 such observations over 12 hours. Two studies have addressed the problem of session length duration in assessment of the activities of mentally retarded people. Van Biervliet (1982) compared data obtained from a smaller number of 42-minute

observation sessions (3 to 19) with a criterion derived from 23 such sessions. Butcher (1983) compared data from up to five 30-minute observation sessions with that obtained from an eight-hour criterion.

In these three studies the representativeness of data taken as the criterion was not assessed. The universe of generalizability was not exhaustively sampled, i.e., the waking hours of a week spent in a residential setting (Van Biervliet, 1982) or the waking hours of a day in an institutional setting (Alevizos et al., 1978; Butcher, 1983). Further, all three studies employed sampling procedures which were not adequately assessed for representativeness in comparison with, say, a continuous real-time record. Thus, the interaction between possible invalidity of sampling procedures (Rojahn & Kanoy, 1985) and non-exhaustive criterion observation sessions can not be ruled out when interpreting their results. What is really required in such studies is a continuous record of behaviours throughout the universe of generalizability as a criterion against which to judge the adequacy of smaller samples.

In the present study sample observation sessions of various durations were computer-simulated from the whole session records described by Study 1. It was hypothesised that any differential representativeness of samples would be related to the basic parameters of the behaviours observed (Tables 4 and 5).

Method

Data analysis

A computer programme (EX1B2, see Appendix 1) was written to

obtain the relative durations of codes in samples extracted from the whole (150-minute) sessions. The starting time (entry point) and duration of the samples were specified to the software by keyboard input. Absolute relative duration of each code in the samples (O) was computed by dividing the number of seconds in which each code was 'on' by the total duration (in seconds) of the sample. The 'real' absolute relative duration (R) of each code in the whole session was calculated and a measure of the percentage similarity (S) between O and R was computed (following Butcher, 1983; Van Biervliet, 1982). S was the result of dividing the smaller of O and R by the larger and multiplying by 100. Finally, S was subtracted from 100 to give a measure of percentage difference between code duration in the whole session and in the sample session. Percentage difference can range between 0%, signifying complete agreement between O and R, and 100%, when R was not zero and O was zero.

Procedure

1) Systematic samples. Nine sample durations were selected ranging from 15 to 135 minutes by 15-minute increments. For each duration three types of samples were taken: around the mid-point of the whole session (see Figure 1); starting at the start of the whole session (see Figure 2); and ending at the end of the whole

---

Insert Figures 1 and 2 about here

---

session. The records of all ten 150-minute sessions were sampled.

Percentage difference was manually recorded from the screen

Figure 1. Samples of increasing duration simulated around the mid-point of sessions.

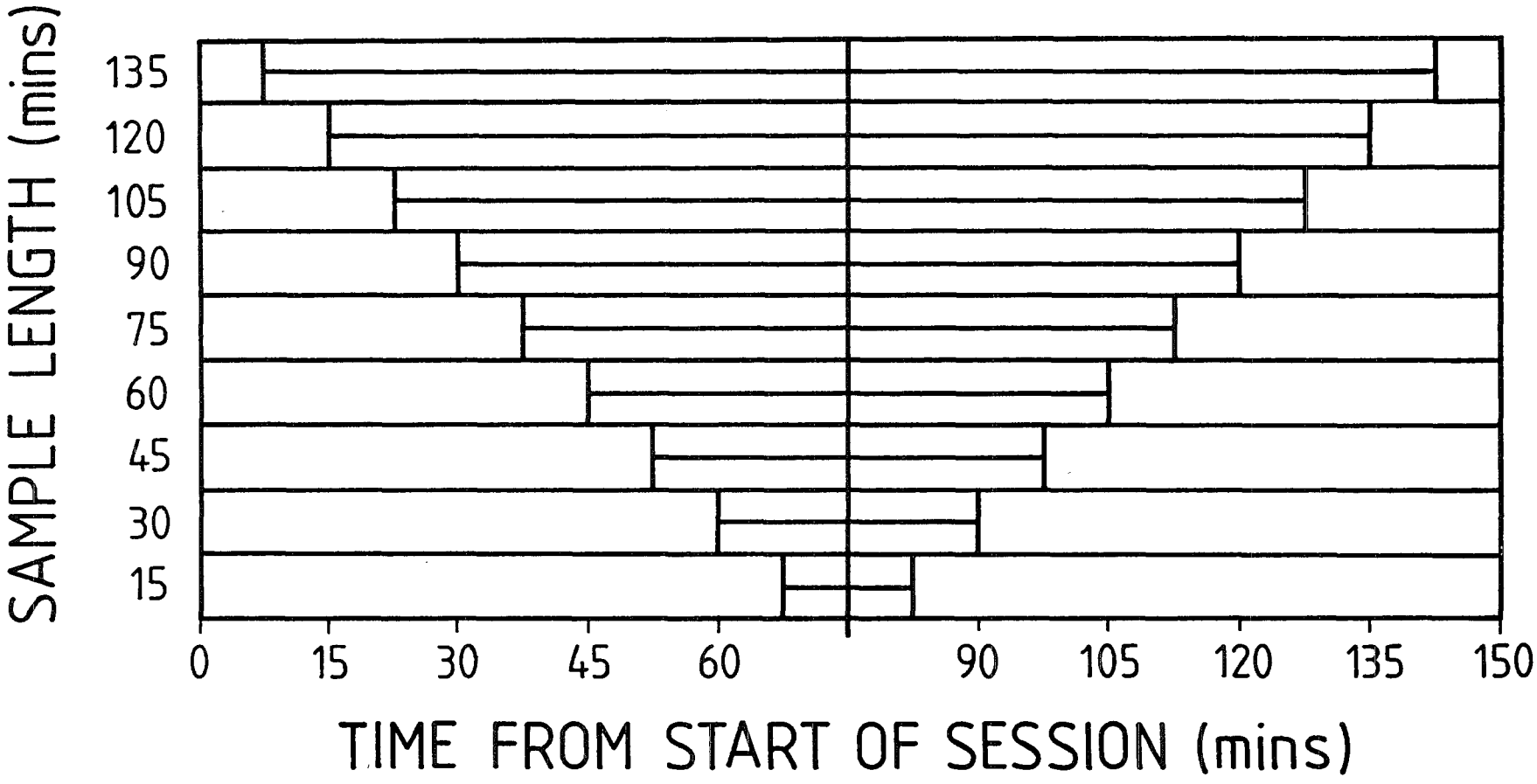
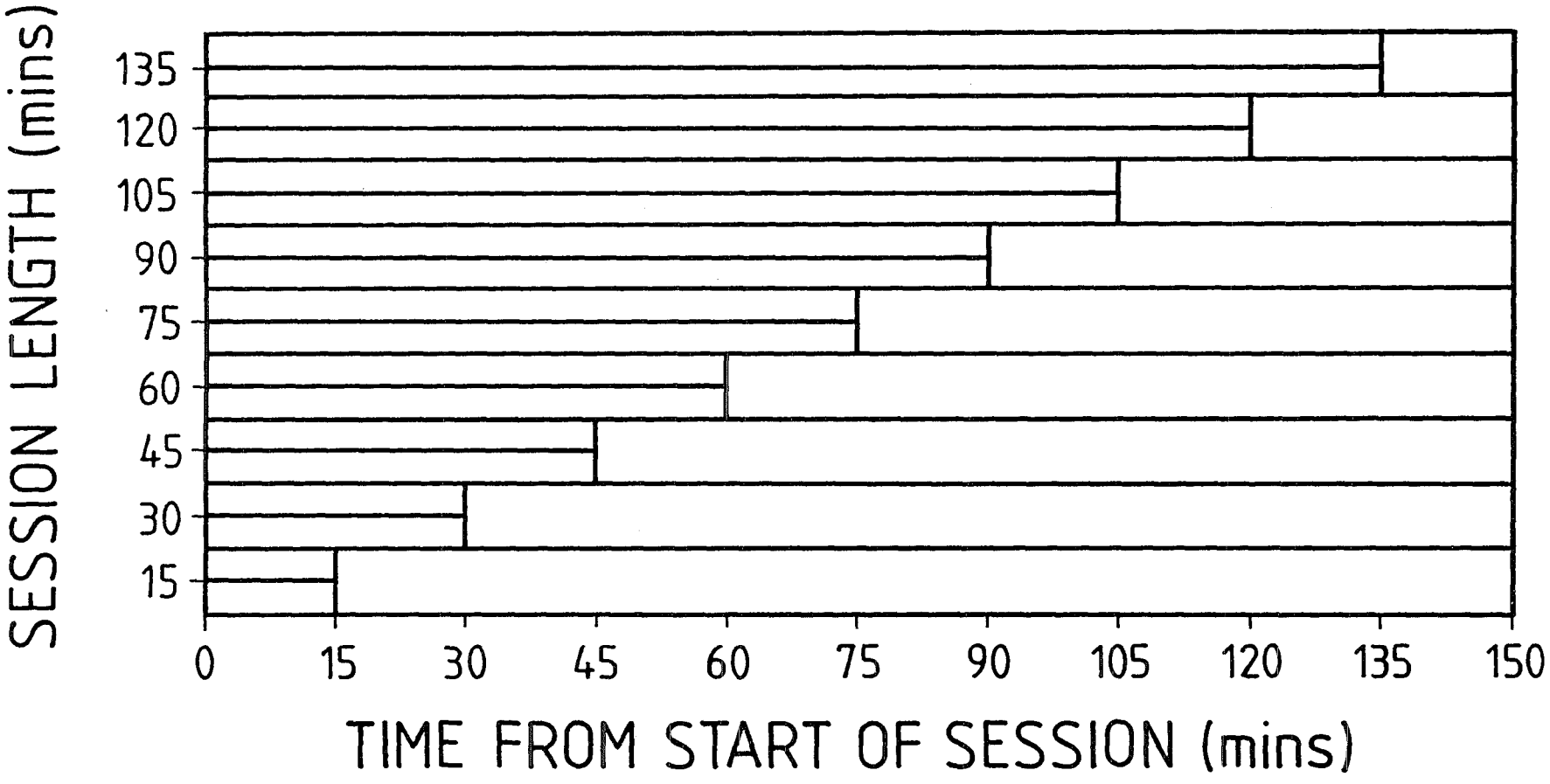

SAMPLE LENGTH (mins)
135
120
105
90
75
60
45
30
15
0
15
30
45
60
90
105
120
135
150
TIME FROM START OF SESSION (mins)

Figure 2. Samples of increasing duration starting at the beginning of sessions (filled bars) and ending at the end of sessions (unfilled bars).


SESSION LENGTH (mins)
135
120
105
90
75
60
45
30
15
0
15
30
45
60
75
90
105
120
135
150
TIME FROM START OF SESSION (mins)

for each code where real relative duration exceeded .007, i.e., the code was on for .7% or more in a session (see Table 4 for real relative duration percentages). Percentage differences for each sample-type were grouped according to values of R for independent behaviour categories (codes 2 to 6) and the mean percentage difference calculated. Groupings of real relative durations were .7 - 2.9%, 3 - 9.9%, 10 - 24.9%, and >50% (as in Table 5). Code 1 was treated separately. The behaviour of staff members, as a group, may have produced different results as some of the parameters in Table 5 for code 1 are dissimilar to the parameters of the 10 - 24.9% grouping into which most sessions' code 1 would fall.

2) Random samples. To ascertain the generality of the findings from the systematic samples a second type of procedure was employed. Random starting times for samples were obtained from random number tables. Sample durations were 15, 45, 105, and 135 minutes. Five random entry points were used for each duration.

Percentage duration was recorded from the screen for codes 2 to 6 and grouped by relative duration, as above. Again, all 10 sessions were sampled. The values of percentage difference became the dependent variable input to analysis of variance (ANOVA, BMDP2V). The random factor was subjects' codes. The grouping factor was the relative duration grouping of the data (three levels) and the within variable was sample duration (five levels).

Results

Systematic samples. Figures 3, 4, and 5 present the mean

Insert Figures 3-5 about here

---

percentage difference (in codes 2 to 6 where relative duration exceeded .007) obtained from samples around the mid-point, starting at the start, and ending at the end of the whole sessions respectively. Figure 6 shows the difference associated with code 1 (social staff).

Insert Figure 6 about here

A general trend towards less difference in the longer duration samples can be seen, with codes of greater relative duration being prone to less difference than more uncommon codes. While these findings support intuition, worthy of particular note is the level of difference, i.e., greater than 20% even with 90-minute samples for codes with relative duration of less than 50%. Codes occurring for more than 50% of a whole session were represented by 30-minute samples to within 20% difference but 15-minute samples were insufficient to achieve this criterion.

Random samples. The means for difference in the random starting-point samples are shown in Figure 7. The effects

---

Insert Figure 7 about here

Figure 3. Percentage difference between relative duration obtained from samples of increasing length and from the whole (150 minute) sessions for subjects' codes grouped by real relative duration. Samples simulated around the mid-point of sessions.

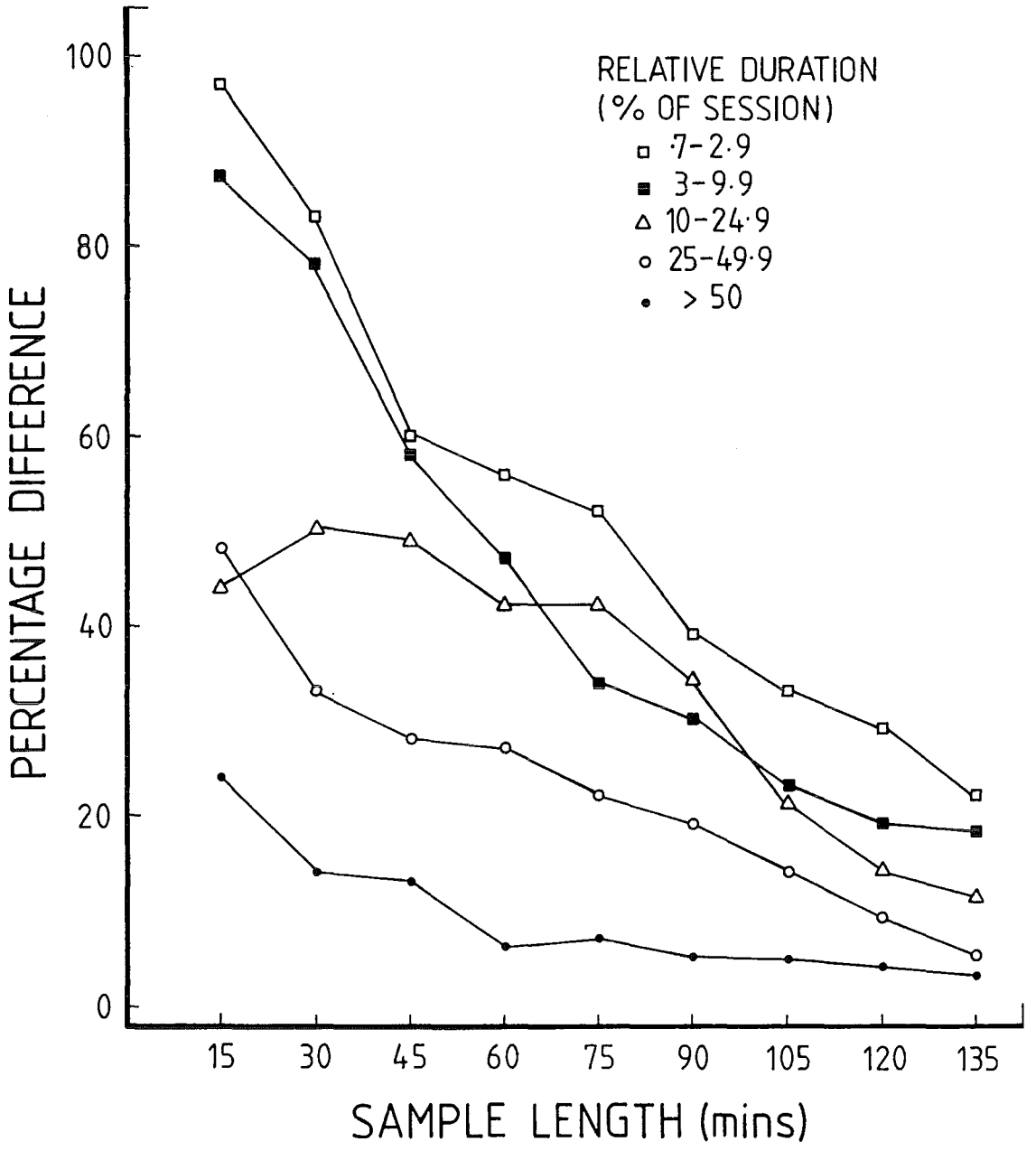

RELATIVE DURATION
(% OF SESSION)
·7-2·9
3-9·9
10-24·9
25-49·9
> 50
PERCENTAGE DIFFERENCE
100
80
60
40
20
0
15
30
45
60
75
90
105
120
135
SAMPLE LENGTH (mins)

Figure 4. Percentage difference between relative duration obtained from samples of increasing length and from the whole (150 minute) sessions for subjects' codes grouped by real relative duration. Samples started at the beginning of sessions.

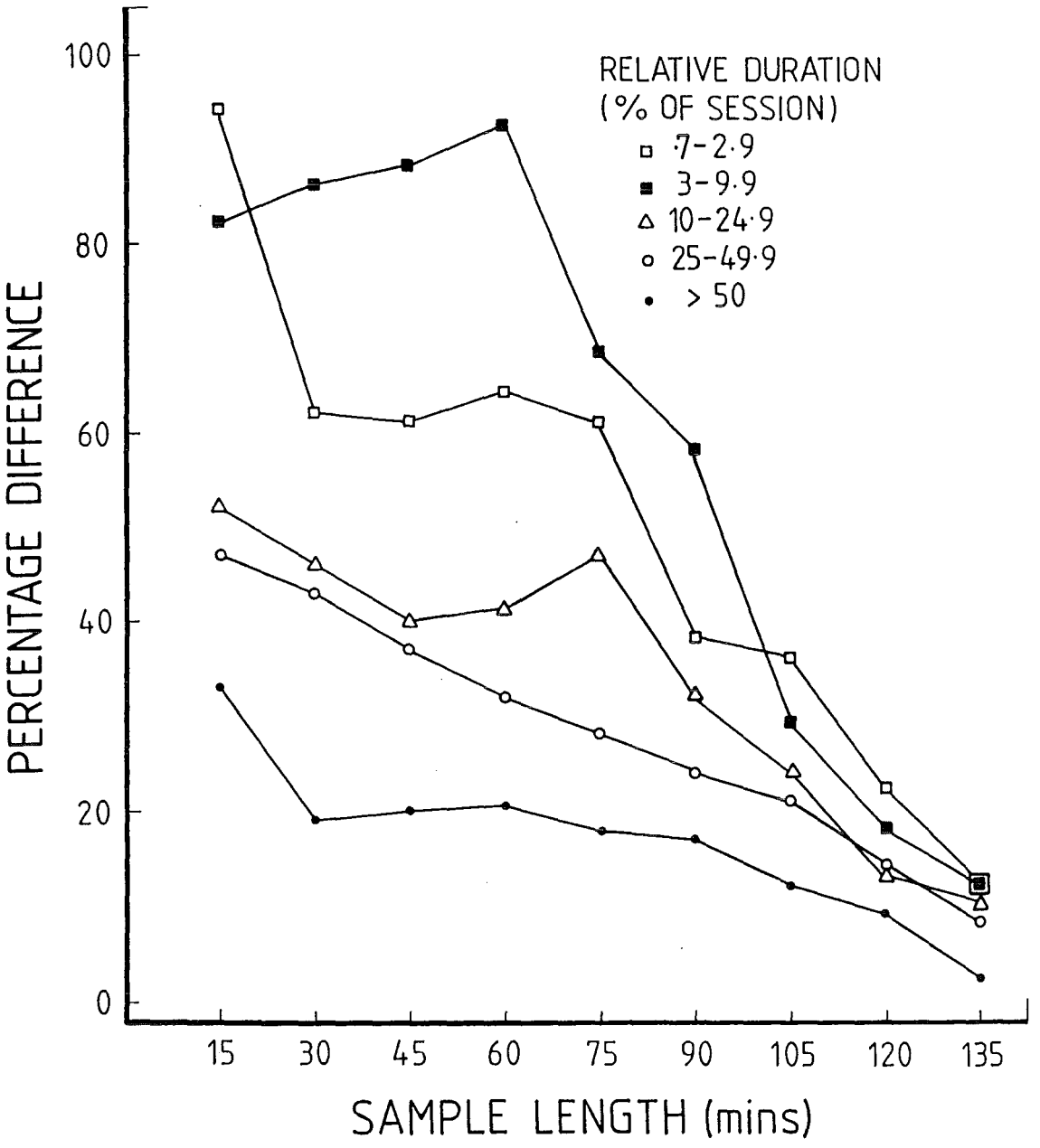

RELATIVE DURATION
(% OF SESSION)
·7-2·9
3-9·9
10-24·9
25-49·9
> 50
PERCENTAGE DIFFERENCE
100
80
60
40
20
0
15
30
45
60
75
90
105
120
135
SAMPLE LENGTH (mins)

Figure 5. Percentage difference between relative duration obtained from samples of increasing length and from the whole (150 minute) sessions for subjects' codes grouped by real relative duration. Samples ending at the end of sessions.

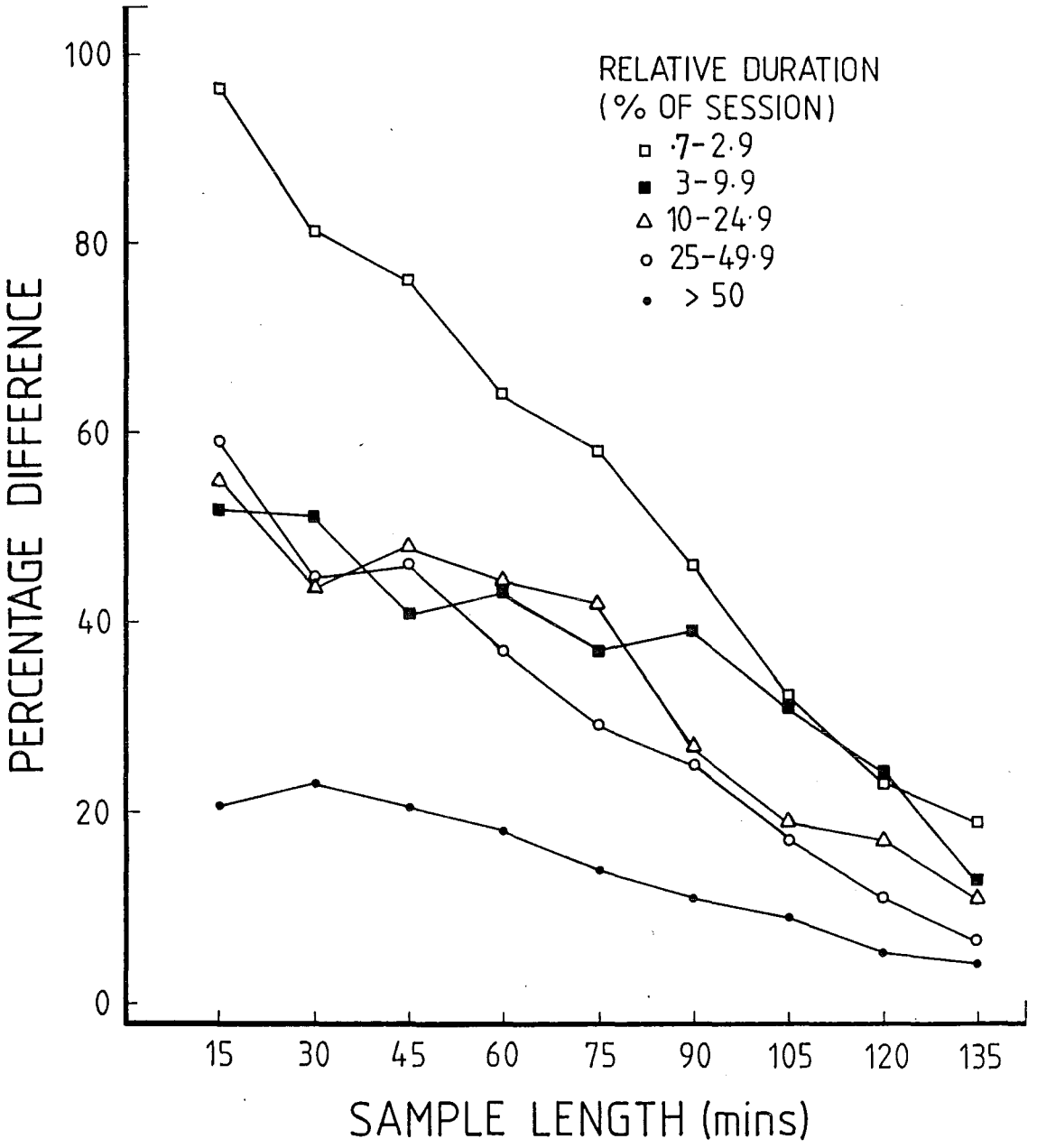

RELATIVE DURATION
(% OF SESSION)
·7-2·9
3-9·9
10-24·9
25-49·9
> 50
PERCENTAGE DIFFERENCE
100
80
60
40
20
0
15
30
45
60
75
90
105
120
135
SAMPLE LENGTH (mins)

Figure 6. Percentage difference between relative durations of code 1 (social staff) obtained from samples of increasing length and from whole (150 minute) sessions. Data grouped according to three systematic sample types.

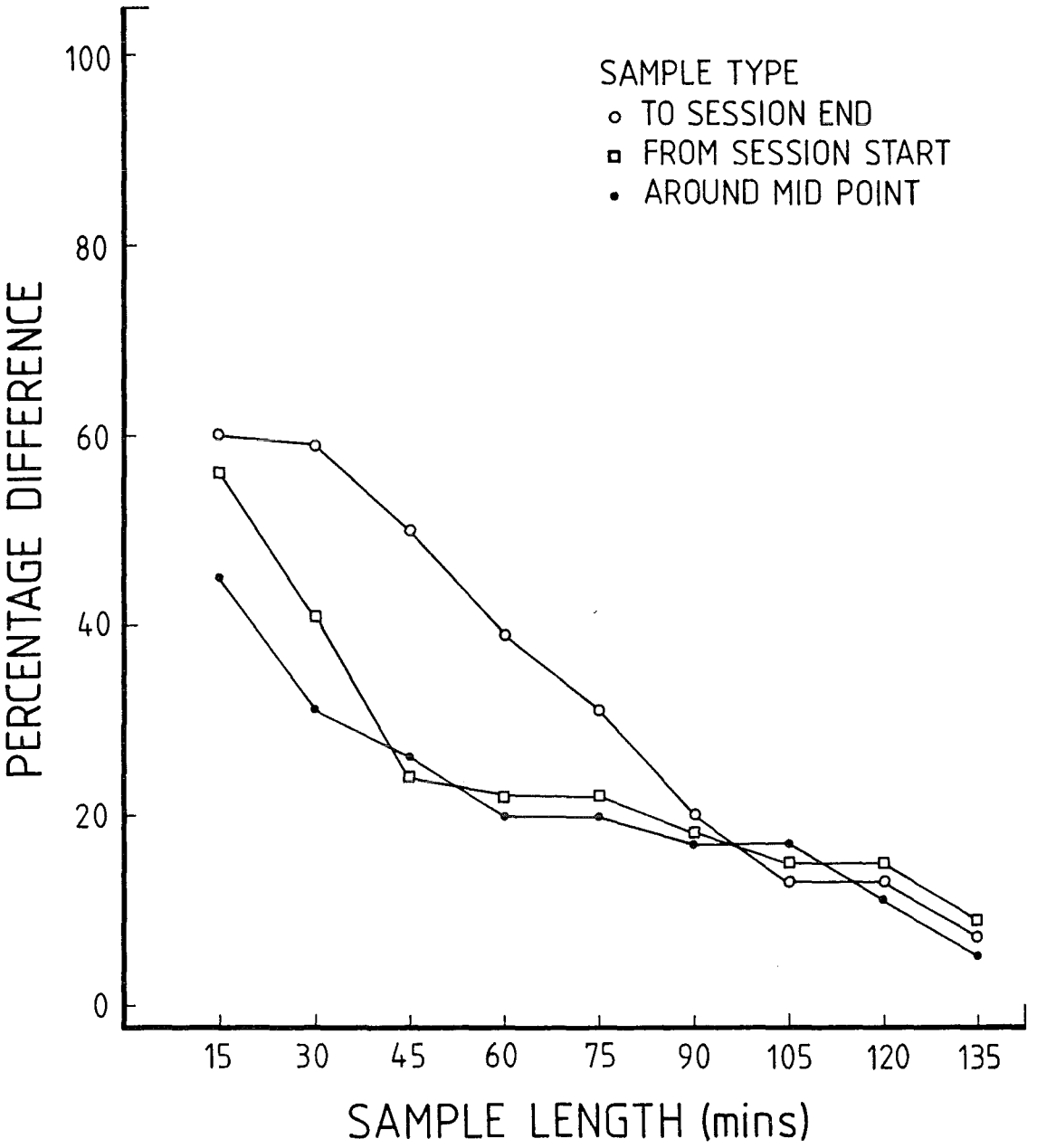
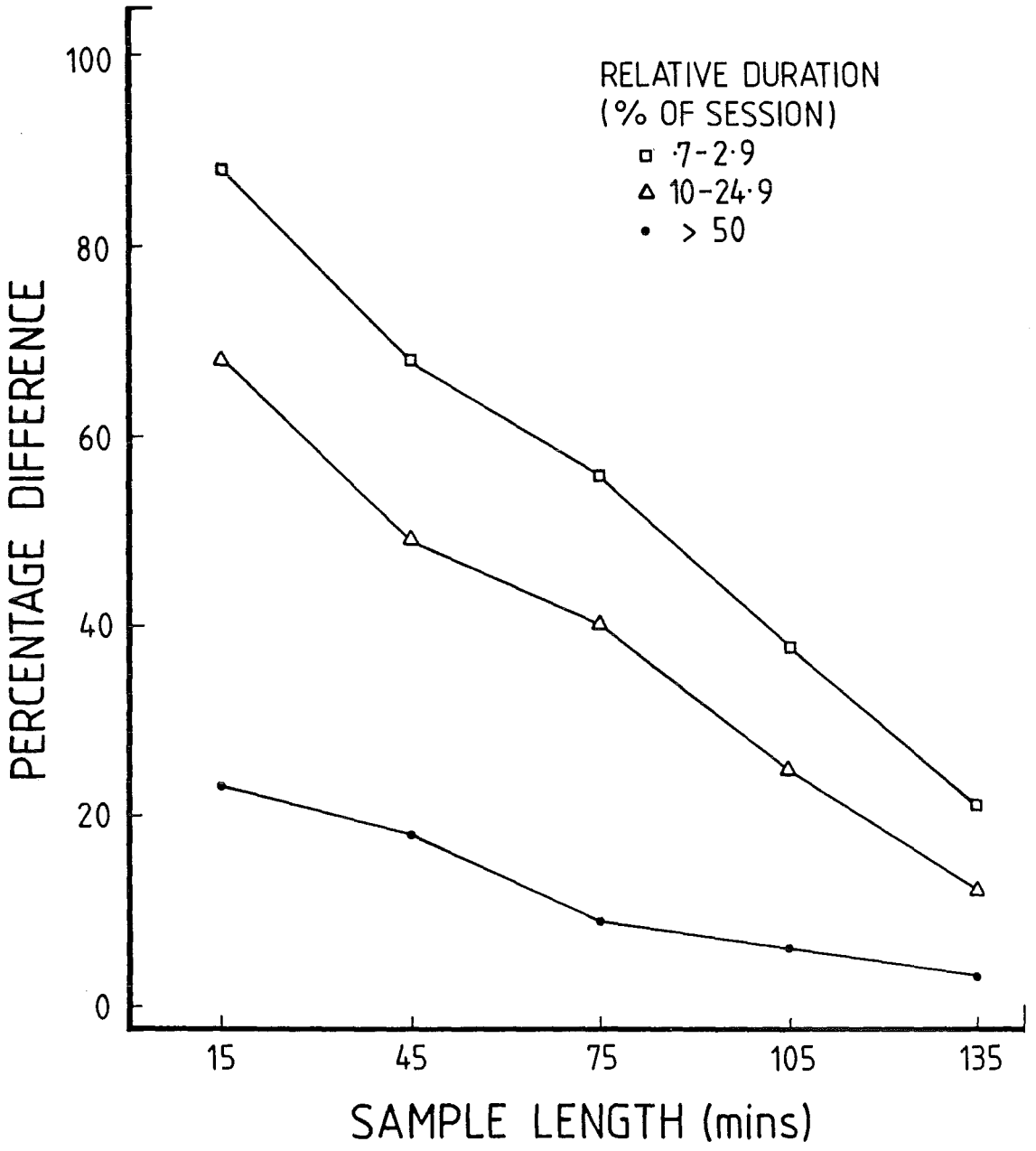

SAMPLE TYPE
TO SESSION END
FROM SESSION START
AROUND MID POINT
PERCENTAGE DIFFERENCE
100
80
60
40
20
0
15
30
45
60
75
90
105
120
135
SAMPLE LENGTH (mins)

Figure 7. Percentage difference between relative duration obtained from samples of increasing length and from the whole (150 minute) sessions for subjects' codes grouped by real relative duration. Samples started at randomly selected entry-points.


RELATIVE DURATION
(% OF SESSION)
·7-2·9
10-24·9
> 50
PERCENTAGE DIFFERENCE
100
80
60
40
20
0
15
45
75
105
135
SAMPLE LENGTH (mins)

demonstrated with the systematic samples can be seen to have been replicated. ANOVA produced highly significant F-ratios ($p < .0001$) for the grouping factor, relative duration in the whole session [$F(2, 92) = 68.23$], and for the within factor, sample duration [$F(4, 368) = 71.65$]. The interaction between relative duration and sample duration was also highly significant [$F(8, 368) = 5.60$ ($p < .0001$)]. Further analysis was not conducted because the effects are clear (Figure 7) and knowledge of exactly where were the significant differences between the means was considered trivial in this case.

Discussion

Results presented indicate that the representativeness of samples taken from continuous exhaustive sessions is a function of the duration of the samples and of the relative duration of behaviours in the sessions. This was confirmed by ANOVA methods. To obtain equal representativeness from behaviours of unequal relative durations longer observational sessions are required for behaviour categories with shorter relative duration.

The degree of non-representativeness of samples considered tolerable by behaviour analysts varies. Van Biervliet (1982) and Butcher (1983) have suggested that 75% similarity, i.e., 25% difference, is a reasonable criterion. Others (Mansell, 1985; Rojahn & Kanoy, 1985) have implied that 10% error may be tolerable. The present study has demonstrated a method by which researchers can empirically determine the representativeness of samples in cases where the whole time of interest can not

reasonably be observed throughout a study. The question of acceptable difference between sample and session measures is one which ought to be judged by consumers of reseach reports. Judgement is only possible if data such as those presented are included in reports.

Explanation for the obtained results can be obtained from Table 5. Moving from highest to lowest relative duration groupings: mean duration of events decreases in an orderly fashion; maximum duration decreases and obtained IRTs increase. If the behaviours of subjects which were clustered are removed, i.e., where expected vs. obtained IRT comparisons are substantial (>.4), maximum IRTs increase. So, for high relative duration behaviours, non-representativeness was the result of long (and variable) event durations and, for low relative duration behaviours, a result of long (and variable) IRTs or clustering of events.

To illustrate the difference between the distributions of high and low relative duration behaviours within a session, two examples have been selected from subject PD's morning session. From a printed record of the raw data from that session the absolute duration of a high relative duration code (code 5) and of a low relative duration code (code 2) was calculated in 15-minute blocks throughout the session. Absolute durations were cumulated across 15-minute periods and divided by the total absolute duration of the codes for the session. This calculation produced a measure of cumulative duration as a proportion of total duration. This measure for these codes was graphed against

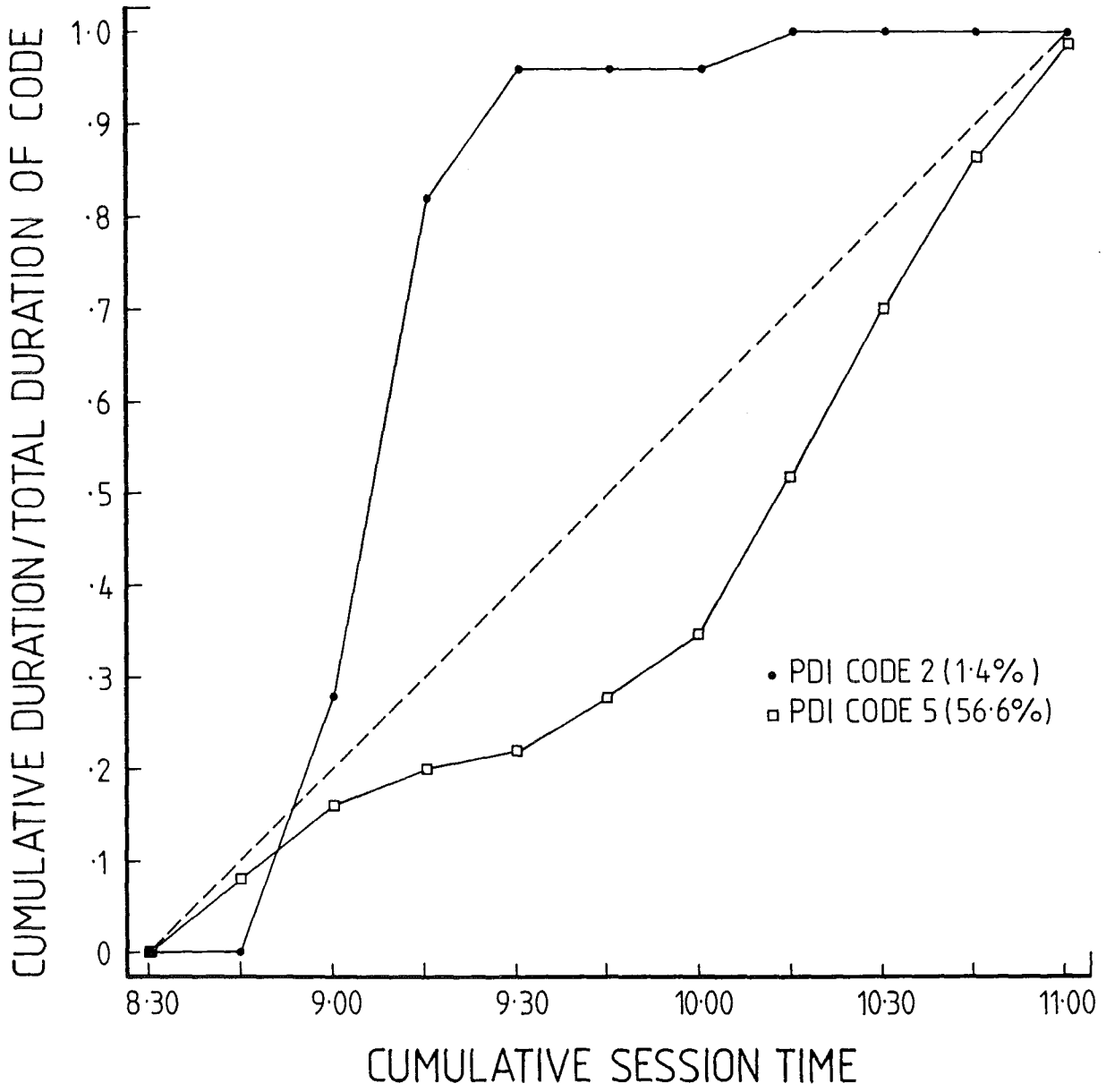
cumulative session duration and is displayed in Figure 8.

Insert Figure 8 about here

Figure 8 can be interpreted as follows: The closer the slope of the line joining points for a code is to the diagonal the greater was the representativenss of shorter duration observational samples. If, in spite of unequal relative durations, behaviours represented by both codes had been emitted at an even rate throughout the session, both lines would be close to the diagonal. Therefore, it is not relative duration per se which has produced the differential representativeness but the distribution of events within the records. In Figure 8, it can be seen that 96% of events coded as social peer (code 2) occurred between 8.45 a.m. and 9.30 a.m. and neither this period, which overestimates relative duration in the whole session, nor the period afterwards, which underestimates it, are representative. Similarly, 65% of hands on objects (code 5) occurred in the last hour of the session.

Butcher (1983) employed sequential level autolag analysis (Sackett, 1978, 1979) to give some explanation for his data on representativeness of observational samples. However, the computation of mean and maximum durations and IRTs seems conceptually more straightforward and provides much the same information as level autolags. The high relative duration behaviours which were measured with Butcher's subjects showed significant autolag correlations, i.e., they occurred in runs. I

Figure 8. Cumulative duration/total duration for a high relative duration code (5) and a low relative duration code (2) from subject PD, morning session. Proportion of code duration recorded plotted against session time.


CUMULATIVE DURATION/TOTAL DURATION OF CODE
1·0
·9
·8
·7
·6
·5
·4
·3
·2
·1
0
8·30
9·00
9·30
10·00
10·30
11·00
CUMULATIVE SESSION TIME
• PDI CODE 2 (1·4%)
□ PDI CODE 5 (56·6%)

suspect that computation of mean event durations would have served the same purpose although the use of dominant category interval recording, rather than real-time, would have precluded the accurate computation of event durations. It must also be stated that Butcher's data were unsuitable for lag analysis for the same reason.

There is a problem with the employment of the percentage similarity statistic in assessment of the validity of observational data. The same can be said for percentage difference, as used in the present study. To explain, I must return to the basic question about generalizability as represented by the present study. How generalizable are data on the relative duration of behaviours in the sample sessions to the relative durations in the whole sessions, or criterion? The question was not: How generalizable are data from the samples to the criterion, or vice versa? The second question is addressed by percentage similarity. The criterion is awarded no special status as it can be either the numerator or denominator in the calculation depending on the value of the measure from the sample (whichever is larger goes on the bottom, i.e., becomes the denominator). Logically, percentage similarity measures do not permit discussion of accuracy, error, representativeness, validity, or generalizability of sample data. This is a fault not acknowledged by previous users of percentage similarity as a dependent variable (Butcher, 1983; Van Biervliet, 1982).

A measure which does allow for evaluation of error in the sample with respect to the criterion would have greater validity,

i.e., 'meaning' in generalizability studies, than percentage difference. A measure of error which has been used in comparisons of sampling methods against a criterion is: Error equals (criterion measure minus sample measure) divided by criterion measure (Mansell, 1985). This can be multiplied by 100 to give a measure of the percentage error in the sample as a proportion of the criterion (Rojahn & Kanoy, 1985). This measure, percentage error, respects the status of the criterion as the sample is judged against it in a consistent manner.

When percentage error is computed, estimates of error can range from +100%, when the sample is zero and the criterion is non-zero, i.e., sample grossly underestimates criterion, to very large negative values when the sample is a gross overestimate. For cases where the sample underestimates the criterion, the values of percentage error and percentage difference are identical. However, percentage difference underestimates error when the sample is an overestimate compared to the criterion. Incidentally, the maximum error possible from a sample overestimate can be calculated. With the present type of data, relative duration, maximum negative error is (relative duration in the session minus 100) divided by relative duration in the session, all times 100. Relative duration in the session can be found in Table 4. The figure of 100 is subtracted as this is the maximum duration, as a percentage, that a code can be on in a sample. This calculation is not quite so easy if the total absolute duration of the code in the session exceeds the sample length. The maximum percentage difference obtainable when the

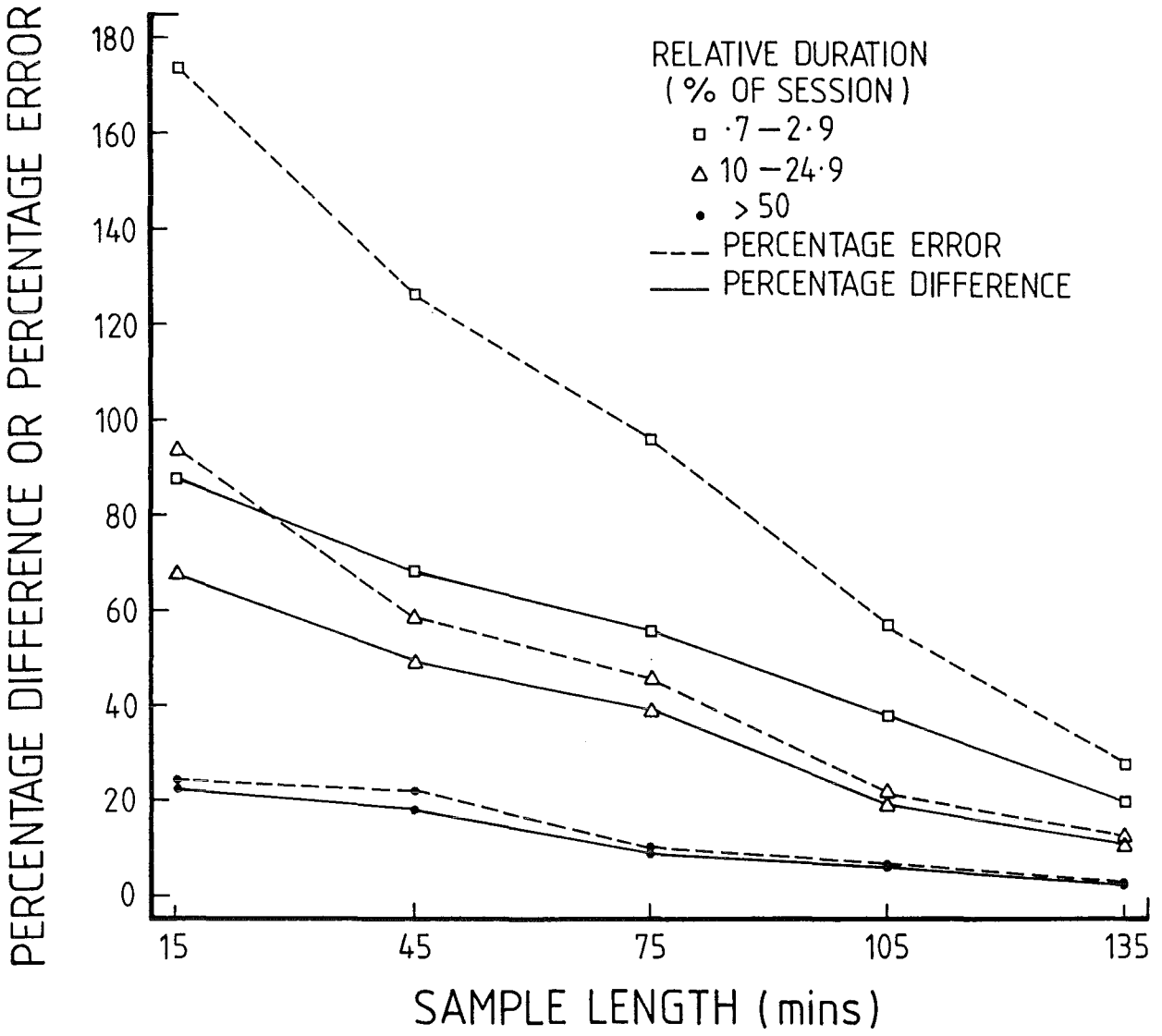
sample overestimates the criterion is equal to 100 minus the relative duration in the session.

To illustrate the effect of employing percentage error as the measure of choice, the same sample observation sessions which were used to assess percentage difference at random entry-points have been re-analysed. A computer programme (ERROR) was written for this analysis (see Appendix 1). The data for percentage difference and percentage error were averaged across the samples starting from five random entry-points at five sample durations and within relative duration groupings. For percentage error the sign of the error was ignored, i.e., 50% overestimates and 50% underestimates were both treated as 50% errors. Figure 9 shows the means obtained by both methods. For percentage difference, the data are identical to that shown in Figure 7.

Insert Figure 9 about here

Two questions arise from examination of Figure 9. How can the increasing differential between error and difference as relative durations decrease be accounted for? What effect does the differential have on interpretation of validity studies when percentage similarity, or difference, has been the statistic selected (Butcher, 1983; Van Biervliet, 1982; and the present study)? Mean percentage error increases dramatically as the probability of at least some gross overestimates from the sample increases. Consider the distribution of event durations in Figure 8. A 15-minute sample starting at 9.03 a.m. overestimated the

Figure 9. A comparison of the percentage difference and percentage error measures for groups of subjects' codes with low, moderate, and high relative durations.


PERCENTAGE DIFFERENCE OR PERCENTAGE ERROR
180
160
140
120
100
80
60
40
20
0
15
45
75
105
135
SAMPLE LENGTH (mins)
RELATIVE DURATION
(% OF SESSION)
·7 — 2·9
10 — 24·9
> 50
PERCENTAGE ERROR
PERCENTAGE DIFFERENCE

relative duration of code 2 by 508% error (but 84% difference). Samples of 15-minute duration starting at 9.37, 9.53, and 10.42 a.m. included no events recorded as code 2; therefore, error was 100%, as was difference. However, the maximum possible negative error for code 5 in this session was 77% (calculated as earlier explained), or (100-56.6)% difference, i.e., 43.4%. Of course, maximum positive error and difference in both cases is 100%.

If, as argued, percentage error is a preferable measure to percentage similarity then conclusions drawn from percentage similarity transformations may need to be re-examined. In the present study, short observation sessions produce more error in estimating relative duration than that indicated by the measure of percentage difference for behaviours occurring for 25% of a session or less. Without data on the basic parameters of behaviours observed in other studies (Butcher, 1983; Van Biervliet, 1982) one can only speculate that the same would be true, i.e., error in sampling low duration behaviours has been underestimated.

Limiting discussion for the present to the subjects, setting, and behaviour categories employed in this study, some recommendations can be made regarding the validity of observational samples. If an assessor of the subjects' performance, e.g., a supervising teacher or psychologist, wished to undertake observations, that person would be well-advised to observe through a whole session. No shorter period is entirely sufficient to obtain a representative sample of all categories of behaviour (see error data for low duration codes in Figure 9).

However, the present data suggest an alternative strategy. Assume that social staff, hands on objects, self-move, and social peer were all ecologically valid categories and desirable, i.e., a social validity study had shown that these are the sort of behaviours that should be displayed in this setting. Given that inappropriate behaviour is rare (i.e., low relative duration), then the assessor need only measure one behaviour category: passive. The desirable categories are measured by the absence of passive provided inappropriate remains at a low level. This strategy can be recommended as passive behaviour had, on average, the highest relative duration (see Study 1). Because of this, sample observations can be relatively free of error when compared to the whole session.

Based on these assumptions, three randomly selected 60-minute sessions from each continuous record were sampled by a variety of computer-simulated momentary time sampling procedures. The computer programme SIMUL was written to simulate various sampling procedures (see Appendix 1). The aim of this was to determine the most economical method for observation of passive behaviour without compromising validity too severely. This represents a partial replication of Mansell (1985) and Sanson-Fisher et al. (1980). Time sampling was chosen because, of all the pencil-and-paper observation procedures, it has been shown to be the most valid in estimating duration (Ary & Suen, 1983; Green et al., 1982; Harrop & Daniels, 1986; Milar & Hawkins, 1976; Murphy & Goodall, 1980; Powell et al., 1977). Also momentary time sampling is probably the least demanding method as the subject

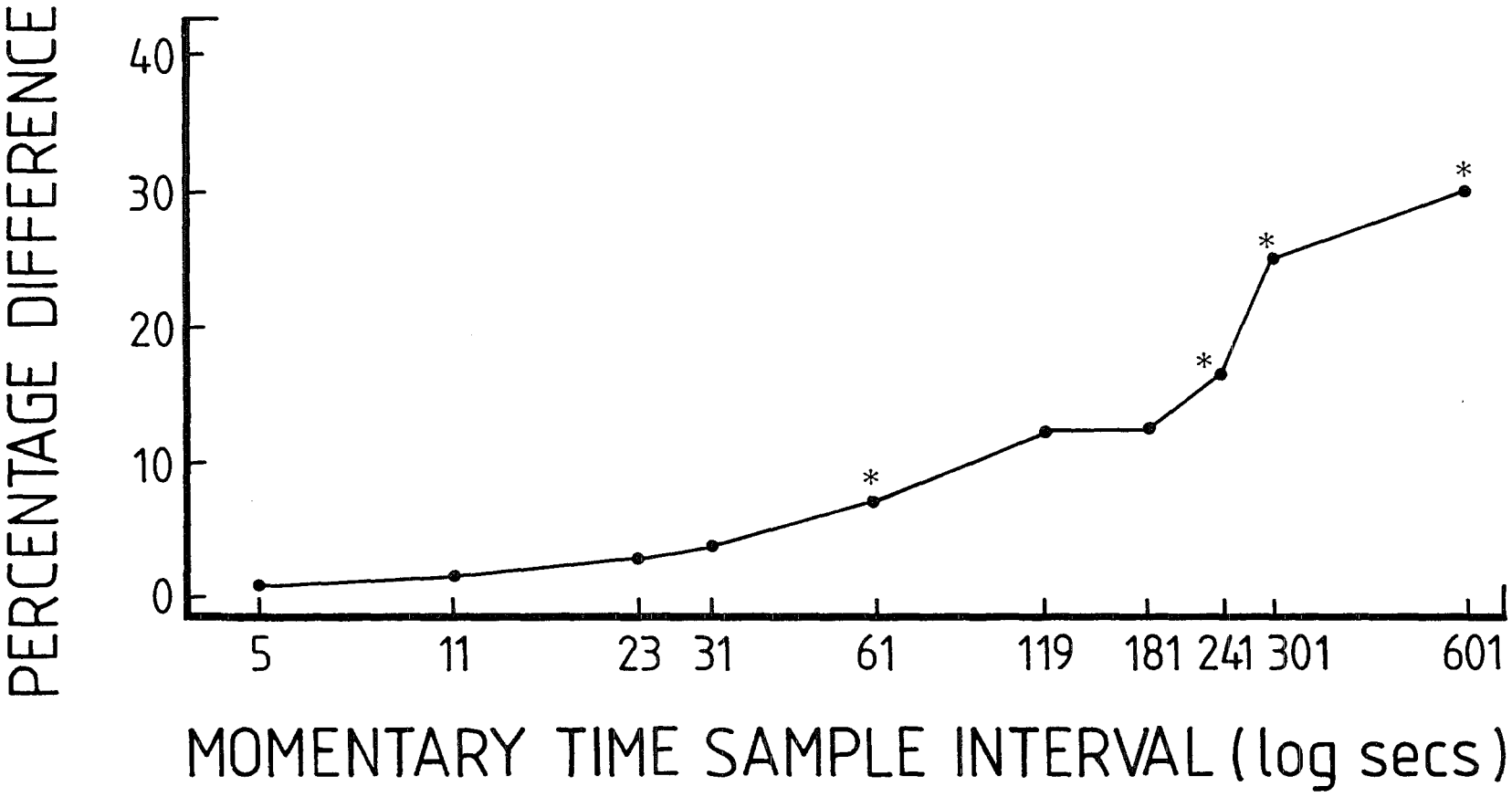
need only be observed at the end of a specified interval. Intervals between samples varied between 5 and 601 seconds. Percentage difference was computed for relative durations of code 6 for each of the simulated observation procedures against the real relative duration within the 60 minutes sampled. Results appear in Figure 10.

Insert Figure 10 about here

If up to 25% difference is taken as an acceptable level of validity, then it can be seen that a time sample every 301 seconds is sufficient to meet this criterion. This is only one sample every five minutes or 12 in one hour. If 20% difference is considered tolerable, one sample every four minutes is sufficient. ANOVA (BMDP2V) was performed on the data and the length of the inter-observation interval was significant [$F(9, 261) = 17.87$, $p < .001$]. The results of a subsequent Tukey HSD test on adjacent means are shown by asterisks in Figure 10. If percentage error had been computed rather than percentage difference, the results would have been similar for a relatively high duration code, as was code 6.

To a person assessing performance, the data on representativeness of samples presented in this paper may make difficulties in sampling seem insurmountable. However, as demonstrated, there is probably a way to avoid days of continuous observation provided that, at least initially and perhaps intermittently, someone performs these observations continuously

Figure 10. Percentage difference between real relative duration and relative duration derived from simulated momentary time samples of increasing inter-sample intervals. Code 6 (passive) was measured for three randomly selected hours in each session. Each data-point represents the mean of 30 measures. Points marked with an asterisk were significantly different from both adjacent points (Tukey HSD = 3.40%, $p < .05$).


PERCENTAGE DIFFERENCE
40
30
20
10
0
5
11
23
31
61
119
181
241
301
601
*
*
*
*
MOMENTARY TIME SAMPLE INTERVAL (log secs)

and in real-time.

Without similar procedures being applied to assess the validity of samples of quantified performance with other subjects, in different settings, and with other behaviour categories, generalizing results beyond the present study is merely speculative. It must be acknowledged that the studies by Butcher (1983) and Van Biervliet (1982), although methodologically flawed, do provide support for the general conclusion that assessment of the validity of samples with respect to the time of interest is neccesary. The neccesity stems from the level of error (or difference), indicating lack of validity, which has been found in the present study, especially as shown in Figure 9.

So far, the present results have been discussed as they might apply to assessors of services, i.e., those seeking an overall picture of a subject's performance. The implications for behaviour therapists who aim to raise or lower the relative durations of target behaviours are somewhat different. If a standard observation/treatment session is employed throughout a study the validity of the data with respect to the whole time of interest is likely to vary across sessions if the intervention is successful. The direction of variation can be predicted. If duration is being reduced then error increases, i.e., validity decreases. The therapist should increase observation session duration to compensate for reduced generalizability. Perusal of the treatment literature as applied to mentally retarded people indicates that this has never been considered. Conversely, as

duration of behaviour increases the validity will increase. Shorter observation sessions may suffice. Here, it seems, is a rationale for economy of time and resources. (However, it is not being suggested that treatment be given for shorter sessions although that may be part of a maintenance package.) As recommended for assessors, therapists will need to evaluate the generalizability of their observation sessions. In doing so by the method used in the present study, the type of data shown in Figure 8 can be produced which could indicate what time of day the target behaviours are most likely to occur. This can provide for a data-based decision about the best time of day to implement intervention.

Finally, this study has provided a demonstration of one of the advantages of real-time recording. From the representation of the stream of behavioural events obtained by this method, the analysis of data is simplified (once the computer programmes have been written) and the basic parameters of behaviour can be measured unambiguously, allowing greater flexibility and detail in analysis. While it has to be admitted that only 2.5 hours observation was sufficient in the present study to exhaustively observe the universe of Time chosen, it would not be practically impossible to schedule observations continuously over longer time periods, i.e., larger universes.

General Discussion

This study has investigated the recognised but insufficiently researched problem of the validity of data obtained from observation samples with durations shorter than the whole time of interest. The generalizability universe (time of interest) was a 2.5-hour training session. It was found that validity, measured either as percentage difference or percentage error, increased as sample durations were increased. In general, and particularly for samples of less than 60 minutes, validity was higher (i.e., difference and error was less) for behaviours with greater relative durations in the session. This result was to be expected; it is intuitively obvious. However, what was surprising was the degree of invalidity with shorter duration samples. If 20% error can be tolerated, it was not until 105-minute samples were taken that this criterion was reached for behaviours occurring for 10-25% of the session. This criterion was not reached, even with 135-minute samples, for low frequency short duration behaviours occupying 1-3% of a session. Samples of 60-minutes duration were sufficient to reflect the relative duration of behaviours taking up more than 50% of a session.

Because raw data were collected by a real-time recording system a fine-grained analysis of the subjects' performance was practicable. It was found that the results could be explained by examining the distribution of behaviours within a session. The more evenly distributed behaviours, which were also of the highest relative duration, were able to be sampled with less

error than the unevenly distributed behaviours.

The question posed by Cone (1977) as to the comparability of data obtained from samples with those obtained from the entire universe of measurement time has been approached. The extent to which behavioural data collected from observational samples of different durations can be generalized to the Times universe sampled depends on the temporal distribution of the behaviours within the universe. The parameters describing temporal distribution can not be known a priori. The recommendation of Johnson and Pennypacker (1980), that exhaustive sampling should precede selection of observational session duration, has been supported empirically by the present findings. Unless the behaviour of interest occurs at a high relative duration (> 50%), there is no support for the recommendation of a standard period of one hour (Bijou et al., 1969; Kazdin, 1984b).

The recommendation that appropriate session length ought to be empirically determined parallels similar advice regarding sampling methods within sessions. Sanson-Fisher et al. (1980) demonstrated how the adequate but economical length of intervals for partial interval recording methods could be determined from a real-time database. The present Study 2 has shown how a similar method can be used to assess the adequacy of session durations. Indeed, as also shown, the same database can be used for both purposes although, to repeat, momentary time sampling has advantages over interval recording if real-time recording can not be routinely employed.

One could speculate that the present findings could have

been obtained from computer-generated records of pseudo-behaviours (e.g., Green & Alverson, 1978; Rojahn & Kanoy, 1985). However, it is difficult to imagine what combination of time series models would generate the data actually obtained from direct observation (Table 5). Rojahn and Kanoy (1985) recommended that an estimation of error should be calculated from computer generated records and tabulated to assist in choice of time sampling parameters. The same suggestion could apply to session length. But, in both cases the basic parameters of the target behaviours (i.e., frequency, duration, and interresponse times) need to be ascertained before tables could be consulted. If data are collected as a real-time record, at least occasionally, then, with the type of software developed for the present study, consultation (even, production) of tables derived from mathematical models is rendered unnecessary. This represents a technical advance which need not be seen as technically pretentious but as empirically justifiable.

Aside from the problems of deciding what behaviours to observe and how to define them adequately, recognised sources of invalidity due to the type of data collection procedures used by behaviour analysts are threefold: interobserver disagreement, recording method, and session length. An important consideration is the effect of the combination of errors produced by these sources. The answer is not simple as invalidity due to observer errors is typically assessed by comparing the performance of observers with one another rather than against a criterion. Invalidity due to recording method and session length is assessed

against a criterion, a real-time continuous whole session record in the present case. Thus, only two of the three sources of error can be combined mathematically unless observer reliability has also been assessed against a criterion. Further, the data presented in Study 2 has been of average error which has disregarded the sign of the error (i.e., over- and underestimates were treated equally) and of average percentage difference, the calculation of which ignores the direction of the difference. As a result it would not be proper to try to demonstrate error combination with sample length data (e.g., Figure 9) and momentary time sample data (Figure 10) because the errors could conceivably cancel or compound either as under- or overestimates. However, as an example, if 66% was obtained as the relative duration from a recording method which produced a 10% overestimate and with a sample length error of 20% overestimate, the 'real' relative duration would have been 50%. The compounded error is a 32% overestimate, a significantly large error from two smaller errors which may be judged individually as acceptable. Observer error against the criterion would again compound upon the total error.

The interaction between sources of invalidity has apparantly received no attention in the literature. It must be emphasised that percentage error from different sources and in the same direction is not merely additive but compounded. This strengthens the forgoing recommendations for the assessment of the validity of observational procedures. Perhaps the compounded error due to recording method and session length could be graphically

displayed as an 'error range' in each phase of intervention studies in a manner similar to the 'disagreement range' for invalidity due to observers (Birkimer & Brown, 1979). This would enable researchers and consumers to judge the adequacy of data collection methods to detect experimental effects.

Behaviour analysts are often reminded that generalization of data beyond the observation sessions to other settings or times of the day can only be empirically justified (e.g., Jones, 1977; Nelson & Hayes, 1979). It can be argued from the present findings and two further premises that data should be presented in applied studies on the degree of empirical generalizability to the whole time of interest. First, applied behaviour analysis has always been concerned with important social problems (Baer, Wolf, & Risley, 1968). Surely it is important to ascertain and report the validity of observational methods, including session length, used to measure the behaviours which constitute these problems. If a problem can occur during, say, a school day, the whole day is important to those involved; not just the half-hour or so of observation. Second, in spite of warnings against it, probably many consumers of research reports do infer generalization across a day beyond the data presented. Such naive beliefs may lead to unjustified expectation of success of experimental programmes at alleviating important social problems. Of concern to assessors of services is that incorrect decisions may be made by administrators on the basis of non-representative samples of performance.

References

Ahrens, M. G. (1986). Psychopaedic training officers. Mental Handicap in New Zealand, 11 (4), 14-43.

Alevizos, P., DeRisi, W., Liberman, R., Eckman, T., & Callahan, E. (1978). The behaviour observation instrument: A method for program evaluation. Journal of Applied Behavior Analysis, 11, 243-257.

Altmann, J. (1974). Observational study of behavior: Sampling methods. Behaviour, 49, 227-267.

Ary, D. (1984). Mathematical explanation of error in duration recording using partial interval, whole interval, and momentary time sampling. Behavioral Assessment, 6, 221-228.

Ary, D. & Suen, H. K. (1983). The use of momentary time sampling to assess both frequency and duration of behavior. Journal of Behavioral Assessment, 5, 143-150.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis, 1, 91-97.

Balsam, P., Fifer, W., Sacks, S. G., Silver, R. (1984). Microcomputers in psychology laboratory courses. Behavior Research Methods, Instrumentation, and Computers, 16, 150-152.

Bates, P. E. & Hanson, H. (1983). Behavioral assessment. In J. L. Matson & S. E. Breuning (Eds.), Assessing the mentally retarded (pp. 27-63). New York: Grune & Stratton.

Bijou, S. W., Peterson, R. F., & Ault, M. H. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. Journal of Applied Behavior Analysis, 1, 175-191.

Bijou, S. W., Peterson, R. F., Harris, F. K., Allen, E., & Johnston, M. S. (1969). Methodology for experimental studies of young children in natural settings. Psychological Record, 19, 143-150.

Birkimer, J. C. & Brown, J. H. (1979). A graphical judgemental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. Journal of Applied Behavior Analysis, 12, 523-533.

Brennan, R. L. & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement, 41, 687-699.

Booth, C. L., Mitchell, S. K., & Solin, F. K. (1979). The generalizability study as a method of assessing intra- and interobserver reliability in observational research. Behavior Research Methods and Instrumentation, 11, 491-494.

Buckley, D. J., Frazer, B. D., & St. Amour, G. (1979). An inexpensive portable printing recorder for behavioral studies. Behavior Research Methods and Instrumentation, 11, 561-563.

Berk, R. A. (1979). Generalizability of behavioural observations: A clarification of interobserver agreement and interobserver reliability. American Journal of Mental Deficiency, 83, 460-472.

Buell, J., Stoddard, P., Harris, F. R., & Baer, D.M. (1968). Collateral social development accompanying reinforcement of outdoor play in a preschool child. Journal of Applied Behavior Analysis, 1, 167-173.

Burgio, L. D., Page, T. J., & Capriotti, R. M. (1985). Clinical behavioral pharmacology: Methods for evaluating medications and contingency management. Journal of Applied Behavior Analysis, 18, 45-59.

Butcher, M. J. (1983). Representativeness of observational data. Unpublished M.A. thesis. University of Auckland.

Cronbach, L. J. (1970). Essentials of psychological testing. New York: Harper & Row.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled agreement or partial credit. Psychological Bulletin, 70, 213-220.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. Behavior Therapy, 8, 411-426.

Deni, R., Szijarto, K., Eisler, A., & Fantauzzo, C. (1983). BASIC programs for observational research using the TRS-80 Model 100 portable and Model 4 computers. Behavior Research Methods and Instrumentation, 15, 616.

Flowers, J. H. (1982). Some simple Apple II software for collection and analysis of observational data. Behavior Research Methods and Instrumentation, 14, 241-249.

Flowers, J. H. & Leger, D. W. (1982). Personal computers and behavioral observation: An introduction. Behavior Research Methods and Instrumentation, 14, 227-230.

Foster, S. L. & Cone, J. D. (1980). Current issues in direct observation. Behavioral Assessment, 2, 313-338.

Gelfand, D. M. & Hartmann, D. P. (1984). Child behavior analysis and therapy (2nd. ed.). New York: Pergamon.

Green, C. W., Reid, D. H., McCarn, J. E., Schepis, M. M., Phillips, J. F., & Parsons, M. B. (1986). Naturalistic observations of classrooms serving severely handicapped persons: Establishing evaluative norms. Applied Research in Mental Retardation, 7, 37-50.

Green, S. B. & Alverson, L. G. (1978). A comparison of indirect measures for long duration behaviors. Journal of Applied Behavior Analysis, 11, 530. (NAPS document # 03288)

Green, S. B., McCoy, J. F., Burns, K. P., & Smith, A. C. (1982). Accuracy of observational data with whole interval, partial interval, and momentary time-sampling recording techniques. Journal of Behavioral Assessment, 4, 103-118.

Grossman, H. J. (Ed.) (1983). Classification in mental retardation. Washington, DC: AAMD.

Goldfried, M. R. (1983). Behavioral assessment. In I. B. Weiner (Ed.), Clinical methods in psychology (2nd ed.) (pp. 233-281). New York: Wiley.

Harmatz, M. G., Mendelsohn, R., & Glassman, M. L. (1975). Gathering naturalistic, objective data on the behavior of schizophrenic patients. Hospital and Community Psychiatry, 26, 83-86.

Harris, F. C. & Lahey, B. B. (1982). Recording system bias in direct observational methodology: A review and critical analysis of factors causing inaccurate coding behavior. Clinical Psychology Review, 2, 539-556.

Harrop, A. & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. Journal of Applied Behavior Analysis, 19, 73-77.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. Journal of Applied Behavior Analysis, 10, 103-116.

Hartmann, D. P. (1982). Assessing the dependability of observational data. In D. P. Hartmann (Ed.), Using observers to study behavior: New directions for methodology of social and behavioral sciences (pp. 51-65). San Fransisco, CA: Jossey-Bass.

Hartmann, D. P. (1984). Assessment strategies. In D. H. Barlow & M. Hersen (Eds.), Single case experimental design: Strategies for studying behavior change (2nd ed.) (pp. 107-139). New York: Pergamon.

Hollenbeck, A. R. (1978). Problems of reliability in observational research. In G. P. Sackett (Ed.), Observing behavior: Vol. II: Data collection and analysis methods (pp. 79-98). Baltimore, MD: University Park Press.

House, A. E., House, B. J., & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. Journal of Behavioral Assessment, 3, 37-57.

Johnston, J. M. & Pennypacker, H. S. (1980). Strategies and tactics of human behavioral research. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jones, R. R. (1977). Conceptual vs. analytic uses of generalizability theory in behavioral assessment. In J. D. Cone & R. P. Hawkins (Eds.), Behavioral assessment: New directions in clinical psychology (pp. 330-343). New York: Brunner/Mazel.

Jones, R. R., Reid, J. R., & Patterson, G. R. (1975). Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment (pp. 42-95). San Fransisco, CA: Jossey-Bass.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. Journal of Applied Behavior Analysis, 10, 141-150.

Kazdin, A. E. (1980). Research design in clinical psychology. New York: Harper & Row.

Kazdin, A. E. (1982). Single case research designs: Methods for clinical and applied settings. New York: OUP.

Kazdin, A. E. (1984a). The scientist practitioner: Research and accountability in applied settings. New York: Pergamon.

Kazdin, A. E. (1984b). Behavior modification in applied settings (3rd ed.). Homewood, IL: Dorsey.

Kelly, M. B. (1977). A review of the observational data-collection and reliability procedures reported in the Journal of Applied Behavior Analysis. Journal of Applied Behavior Analysis, 10, 99-101.

Klesges, R. C., Woolfrey, J., & Vollmer J. (1985). An evaluation of the reliability of time sampling versus continuous data collection. Journal of Behaviour Therapy and Experimental Psychiatry, 16, 303-307

Koontz, F. W. (1982). WATCH: Microcomputer programs to collect and analyse behavioral observations based on focal-animal sampling. Behavior Research Methods and Instrumentation, 14, 431-432.

Landesman-Dwyer, S. & Sackett, G. P. (1978). Behavioral changes in nonambulatory, profoundly mentally retarded individuals. In C. E. Meyers (Ed.), Quality of life in severely and profoundly retarded people: Research foundations for improvement (pp. 55-144). Washington, DC: AAMD.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Linscheid, T. R., Feiner, J., & Sostek, A. M. (1984). Use of time-lapse video recording for the direct measurement of behavior in the mentally retarded. Applied Research in Mental Retardation, 5, 317-327.

Lovaas, O. I., Freitag, G., Gold, V. J., & Kassorla, I. C. (1965). Experimental studies in childhood schizophrenia: Analysis of self-destructive behaviour. Journal of Experimental Child Psychology, 2, 67-84.

Maclean, W. E. (Jr.), Tapp, J. T., & Johnson, W. T. (1985). Alternate methods and software for calculating interobserver agreement for continuous observation data. Journal of Psychopathology and Behavioral Assessment, 7, 65-73.

Magyar, R. L. & Fitzsimmons, J. R. (1979). A multichannel, portable, "real time", event encoder-decoder for laboratory and field experimenters. Behavior Research Methods and Instrumentation, 11, 47-50.

Mansell, J. (1985). Time sampling and measurement error: The effect of interval length and sampling pattern. Journal of Behaviour Therapy and Experimental Psychiatry, 16, 245-251.

Milar, C. R. & Hawkins, R. P. (1976). Distorted results from the use of interval recording procedures. In T. A. Brigham, R. Hawkins, J. W. Scott, & T. F. McLaughlin (Eds.). Behavior Analysis in Education (pp. 261-273). Dubuque, IO: Kendall/Hunt.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376-390.

Moss, S. C. (1984). On-line Apple II for recording and analysis of complex motor movements. Behavior Research Methods, Instrumentation, and Computers, 16, 19-24.

Mudford, O. C. (1985). Training officers and applied behaviour analysis. Mental Handicap in New Zealand, 9(3), 47-61.

Murphy, G. & Goodall, E. (1980). Measurement error in direct observations: A comparison of common recording methods. Behaviour Research and Therapy, 18, 147-150.

Nelson, R. O. & Hayes, S. C. (1979). The nature of behavioral assessment: A commentary. Journal of Applied Behavior Analysis, 12, 491-500.

New Zealand Psychological Society (1986). Members' Handbook. Auckland, NZ: NZPsS.

Odom, S. L., Hoyson, M., Jamieson, B., & Strain, P. S. (1985). Increasing handicapped preschoolers' peer social interactions: Cross-setting and component analysis. Journal of Applied Behavior Analysis, 18, 3-16.

Parsonson, B. S. & Baer, D. M. (1978). The analysis and presentation of graphic data. In, T. R. Kratochwill (Ed.), Single subject research: Strategies for evaluating change (pp. 101-165). New York: Academic Press.

Pfadt, A. & Tryon, W. W. (1983). Issues in the selection and use of mechanical transducers to directly measure motor activity in clinical settings. Applied Research in Mental Retardation, 4, 251-270

Poole, A. D., Sanson-Fisher, R. W., & Thompson, V. (1981). Observations on the behaviour of patients in a state mental hospital and a general hospital psychiatric unit: A comparative study. Behaviour Research and Therapy, 19, 125-134.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sampling measures of behavior. Journal of Applied Behavior Analysis, 8, 463-469.

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. Journal of Applied Behavior Analysis, 10, 325-332.

Repp, A. C. & Barton, L. E. (1980). Naturalistic observations of institutionalized retarded persons: A comparison of licensure decisions and behavioral observations. Journal of Applied Behavior Analysis, 13, 333-341.

Rojahn, J. & Kanoy, R. C. (1985). Toward an empirically based parameter selection for time-sampling observation systems. Journal of Psychopathology and Behavioral Assessment, 7, 99-120.

Sackett, G. P. (1978). Measurement in observational research. In G. P. Sackett (Ed.). Observing behavior: Vol II: Data collection and analysis methods (pp. 25-43). Baltimore, MD: University Park Press.

Sackett, G. P. (1979). The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In J. D. Osofsky (Ed.). Handbook of infant development (pp. 623-649). New York: Wiley.

Sackett, G. P. & Landesman-Dwyer, S. (1977). Toward an ethology of mental retardation: Quantitative behavioral observation in residential settings. In P. Mittler (Ed.), Research to practice in mental retardation (vol II) (pp. 27-37). Baltimore, MD: University Park Press.

Sackett, G. P., Stephenson, E., & Ruppenthal, G. C. (1973). Digital data acquisition systems for observing behavior in laboratory and field settings. Behavior Research Methods and Instrumentation, 5, 344-348.

Sanson-Fisher, R. W., Poole, A. D., & Dunn, J. (1980). An empirical method for determining an appropriate interval length for recording behavior. Journal of Applied Behavior Analysis, 13, 493-500.

Sanson-Fisher, R. W., Poole, A. D., Small, G. R., & Fleming, I. R. (1979). Data acquisition in real time: An improved system for naturalistic observations. Behavior Therapy, 10, 543-554.

Schinke, S. P. & Wong, S. E. (1977). Coding group home behavior with a continuous realtime recording device. Behavioral Engineering, 4, 5-9.

Simpson, M. J. A. (1979). Problems of recording behavioral data by keyboard. In S. J. Suomi & G. R. Stephenson (Eds.), Social interaction analysis: Methodological issues (pp. 137-156). Madison, WI: University of Wisconsin Press.

Stokes T. F. & Baer, D. M. (1977). An implicit technology of generalization. Journal of Applied Behavior Analysis, 10, 349-367.

Strossen, R. J., Coates, T. J., & Thoresen, C. E. (1979). Extending generalizability theory to single-subject designs. Behavior Therapy, 10, 606-614.

Sulzer-Azaroff, B. & Mayer, G. R. (1977). Applying behavior analysis with children and youth. New York: Holt, Rinehart, & Winston.

Torgerson, L. (1977). Datamyte 9000. Behavior Research Methods and Instrumentation, 9, 405-406.

Towns, A. J., Singh, N. N., & Beale, I. L. (1984). Reliability of observations in a double- and single-blind drug study. In K. Gadow (Ed.), Advances in learning and behavioral disabilities (Vol. 3)(pp. 215-240). Greenwich, CT: JAI Press.

Van Biervliet, A. (1982). Empirical determination of three parameters of an observation system designed for descriptive evaluation of residential services. Unpublished manuscript: University of Otago.

Voeltz, L. M. & Evans, I. M. (1982). The assessment of behavioral interrelationships in child behavior therapy. Behavioral Assessment, 4, 131-165.

Wildman, B. G. & Erickson, M. T. (1977). Methodological problems in behavioral observation. In J. D. Cone & R. P. Hawkins (Eds.), Behavioral assessment (pp. 255-273). New York: Brunner/Mazel.

White, R. E. C. (1971). WRATS: A computer compatible system for automatically recording and transcribing behavioural data. Behaviour, 40, 135-148.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. Journal of Applied Behavior Analysis, 11, 203-214.

Ysseldyke, J. E., Thurlow, M. L., Mecklenburg, C., & Graden, J. (1984). Opportunity to learn for regular and special education students during reading instruction. Remedial and Special Education (RASE), 5, 29-37.

Appendix 1: contents

```
20 DEFSTR A-G: DEFINT U-Z: DEFSNG T: DIM T(5000): DIM A1(5000): DIM Y(5000): DIM
 D(40): U2=1
25 CLS:INPUT"Name file , please -- use test as name for playing";G:G1="B:"+G
26 PRINT"Ensure number lock on for first obs'r";:INPUT" try it";G7:PRINT"Second
obs'r may start any time after first has entered something":PRINT"2nd. obs'r use
s letters a,s,d,z,x,c":PRINT"& signals end for 2nd. obs'r ! = end of session"
27 GOSUB 7000
50 PRINT "Enter code"
200 A=INKEY$: IF A="" THEN 200
201 IF A="t" THEN PRINT TIME$: GOTO 50
204 ON ERROR GOTO 0
205 ON ERROR GOTO 6000
210 IF A=A1(U-1) THEN GOTO 200: REM disallows one code to follow itself
211 ON ERROR GOTO 0
215 V=ASC(A): IF V > 48 AND V < 55 THEN GOTO 220
216 IF V=97 OR V=99 OR V=100 OR V=115 OR V=120 OR V=122 OR V=38 OR V=33 THEN GOT
O 220
217 GOTO 200
220 B=TIME$
300 A1(U)=LEFT$(A,1):PRINT A1(U)
310 T(U)=((VAL(LEFT$(B,2)))*3600)+((VAL(MID$(B,4,2)))*60)+VAL(RIGHT$(B,2)):IF U=
0 THEN T1=T(0)
460 IF A="!" THEN PRINT"End of session":GOSUB 9000:CLS:PRINT"Turn number lock of
f":END
500 A="": U=U+1: GOTO 50
6000 IF U=0 THEN RESUME 211
6010 RESUME 211
7000 CLS:PRINT"These are the instructions for observers using this IBM PC comput
er":PRINT:PRINT"The person sitting on the right is the primary observer: The rel
iability checkeris on the left"
7010 PRINT"The primary observer inputs codes for behaviours by pressing one of t
he numeric keys on the right of the keyboard. The number you press will show on
the screen.The primary observer must press the 'return' key when the tutor tells
 the"
7020 PRINT"observers to start recording and then the number code for the behavio
ur first":PRINT"observed. Only when you see a change in the subject's behaviour
should you inputanother code. When told that the session is finished press the a
rrow pointing
7030 PRINT"up and 1 on the main keyboard to get the ! (exclamation mark).":PRINT
"The second observer must press the letter for the code for the first behaviour
 observed soon after the first observer has input his/her first code and a"
7040 PRINT"different code when you observe a change in behaviour. When the tutor
 tells 2nd.observers to stop you must press the up arrow and 7 on the main keybo
ard to get the & which signals end of reliability check. The 1st. observer conti
nues to "
7043 PRINT"input until the end of the session.":PRINT"Please do not hold the key
s down, just one press is enough.":PRINT"Ignore the blips on the video-tape as t
hey are there for the manual recorders."
7050 INPUT"Press the return key when told to";A:CLS:RETURN
9000 PRINT"Storing data -- DON'T TOUCH ANYTHING": FOR X=1 TO U: Y(X)=T(X)-T(X-1)
: NEXT X: Y(0)=0: X=0
10000 OPEN "R",1,G1: FOR V= 0 TO 38: FIELD 1, (V*3) AS AX, 3 AS D(V): NEXT V: FI
ELD 1, 117 AS AX, 2 AS D1, 8 AS B2
10020 FOR V=0 TO 38: LSET D(V)= A1(X)+MKI$(Y(X)): X=X+1: IF X=U+1 THEN GOTO 1006
0
10050 NEXT V
10060 LSET D1=MID$(DATE$,4,2): LSET B2=MKS$(T1)
10100 PUT 1, U2: U2=U2+1:IF X=U+1 THEN GOTO 10200 ELSE GOTO 10020
10200 CLOSE: RETURN
```

```
10 REM This programme calculates overall kappa (Hollenbeck, 1978)
20 DEFSTR A-G: DEFINT L-O: DEFINT U-Z: DEFSNG T: DIM T(2000): DIM A1(2000): DIM
D(40): U2=1: DIM L(10000): DIM M(10000)
50 INPUT"Name file for reliability check";G:G1="B:"+G
60 GOSUB 10000:V3=0
100 FOR V1=V3 TO (W-1): IF ASC(A1(V1))>96 AND ASC(A1(V1))<123 THEN 150
110 NEXT V1: PRINT "no reliability checks found": END
150 N=0: Y=VAL(A1(V1-1)): GOSUB 1000: T=T(V1): T1=T
200 L(N)=Y: M(N)=X: IF T(V1)<>T THEN N=N+1: T=T+1: GOTO 200
205 IF A1(V1)="&" THEN GOTO 300
210 IF ASC(A1(V1))>96 AND ASC(A1(V1))<123 THEN GOSUB 1000 ELSE Y=VAL(A1(V1))
220 V1=V1+1: IF T(V1)=T THEN 210
230 GOTO 200
250 FOR U=0 TO N: PRINT L(U),M(U),U:STOP:REM checks L and M -- delete later
251 NEXT U
300 T2=T(V1): FOR U=0 TO N: Z(L(U),M(U))=Z(L(U),M(U))+1:NEXT U
310 FOR U=1 TO 9: FOR N=1 TO 9: Z=Z+Z(N,U): X(N)=X(N)+Z(N,U): Y(U)=Y(U)+Z(N,U):
NEXT N: NEXT U
315 CLS:PRINT"Observer agreement matrix table -- primary codes across the top.(
";G;")."
320 FOR N=1 TO 6: PRINT TAB(((N-1)*6)+7) N;: NEXT N:PRINT TAB(58) "tot     p":P
INT: FOR U=1 TO 6:PRINT U TAB(7);: FOR N=1 TO 6: PRINT Z(N,U) TAB((N*6)+7);: NE
T N: PRINT TAB(57); Y(U) TAB(67);:PRINT USING "#.###";Y(U)/Z:NEXT U:PRINT
330 PRINT "tot.";:PRINT TAB(7);:FOR U=1 TO 6: PRINT X(U) TAB((U*6)+7);:NEXT U:P
INT:PRINT "p";:PRINT TAB(8);: FOR U=1 TO 6:PRINT USING "#.###"; X(U)/Z;:PRINT T
B((U*6)+8);: NEXT U:PRINT:PRINT
350 FOR U=1 TO 9: Z1=Z1+Z(U,U): NEXT U: PRINT"z1 = ";Z1;"  z1/z or p0 = ";Z1/Z;
360 FOR U=1 TO 9:T3=T3+(X(U)/Z)*(Y(U)/Z): NEXT U : PRINT "     pc = ";T3:PRINT
362 T4=((Z1/Z)-T3)/(1-T3): T5=T4/(SQR(T3/((T-T1)*(1-T3))))
365 PRINT "KAPPA = ";:PRINT USING "#.##";T4;: PRINT"   Z =";T5;: IF T5>1.96 THE
 PRINT"* at .05" ELSE PRINT "non-significant"
370 PRINT: PRINT" 2nd. obs'r started at ";T1;:PRINT" and finished at ";T;:PRINT
elapsed time = ";T-T1;" secs.":INPUT"Shall I check for more reliability checks
n this file (y/n)";G7
371 IF G7="y" THEN V3=V1:ERASE Z,X,Y:Z=0:Z1=0:T3=0:GOTO 100
380 PRINT" end of reliability calculation": END
1000 IF A1(V1)="z" THEN X=1: RETURN
1010 IF A1(V1)="x" THEN X=2: RETURN
1020 IF A1(V1)="c" THEN X=3: RETURN
1030 IF A1(V1)="a" THEN X=4: RETURN
1040 IF A1(V1)="s" THEN X=5: RETURN
1050 IF A1(V1)="d" THEN X=6: RETURN
1060 IF A1(V1)="q" THEN X=7: RETURN
1070 IF A1(V1)="w" THEN X=8: RETURN
1080 IF A1(V1)="e" THEN X=9: RETURN
1090 STOP
5010 FOR 09 = V1-1 TO 0 STEP -1: IF (ASC(A1(09)) > 48) AND (ASC(A1(09)) <  58)
HEN A1(V1)= A1(09): RETURN
10000 OPEN "R",1,G1: FOR V= 0 TO 38: FIELD 1, (V*3) AS AX, 3 AS D(V): NEXT V: F
ELD 1, 117 AS AX, 2 AS D1, 8 AS B2
10050 GET 1,1: T=CVS(B2)
10100 GET 1, U2:FOR V=0 TO 38: A1(W) = LEFT$(D(V),1): T(W) =CVI(RIGHT$(D(V),2))
T: T=T(W): IF A1(W)= "!" THEN PRINT W;" events input": CLOSE: RETURN
10200 W=W+1: NEXT V: U2=U2+1: GOTO 10100
```

```
10 REM this programme takes varying sessions lenghts and starting times to give
%age occurrence and difference scores -- called EX1b2
20 DEFSTR A-G: DEFINT L-O: DEFINT U-Z: DEFSNG T: DIM TIME(1000): DIM EVENT(1000)
: DIM D(40): U2=1: DIM L(10800):DIM NTOT(1800)
50 INPUT"Name file for analysis of sampling procedures";G:G1="B:"+G
60 GOSUB 10000
100 FOR X=0 TO W: Z=TIME(X)-TIME(0): L(Z)=VAL(EVENT(X)):NEXT X
150 FOR X=0 TO TIME(W)-TIME(0): IF L(X)=0 THEN L(X)=L(X-1)
160 NEXT X
200 ERASE EVENT: ERASE D:REM input ended a.o.k. up to here l(0 to w) is list of
behs. in each sec. from start of session
230 TIMEALL=TIME(W)-TIME(0)
240 FOR X=0 TO TIMEALL-1: N1(L(X))=N1(L(X))+1: NEXT X: FOR X=1 TO 9: T2(X)=(N1(X
)/TIMEALL)*100:NEXT X: REM t2(1-9) are real %ages of occurance
247 N(1)=1:ERASE N:GOSUB 4000
250  N(1)=1:ERASE N:PRINT"select sampling strategy --- enter 1 for momentary tim
e sample"
255 PRINT"                                enter 2 for interval obs'n method":INPUT
M
257 IF M=2 THEN GOTO 1260
280 INPUT"Enter time in secs. between observations";O3:REM point sample starts h
ere
290 NUMOBS=INT(TIMEALL/O3)
350 FOR X=O3-1 TO (NUMOBS*O3)-1 STEP O3:  N(L(X))=N(L(X))+1:NEXT X:C1="points"
360 PRINT" behaviour code      %age of ";C1;"      REAL %age    %age difference"
:FOR X=1 TO 6: PRINT"  ";X;"                    ";(N(X)/NUMOBS)*100;"          ";:PR
INT TAB(42) T2(X);: REM for 9 codes change 6 to 9
363 IF T2(X)=0 AND N(X)=0 THEN PRINT TAB(56) " 100 equal":GOTO 369
364 IF T2(X)=0 OR N(X)=0 THEN PRINT TAB(56) " zero ":GOTO 369
365 IF T2(X)/((N(X)/NUMOBS)*100)=1 THEN PRINT TAB(56)" 100 equal":GOTO 369
366 IF T2(X)>(N(X)/NUMOBS)*100 THEN PRINT TAB(56) 100-((((N(X)/NUMOBS)*100)/T2(X
))*100);" under":GOTO 369
367 PRINT TAB(56) 100-((T2(X)/((N(X)/NUMOBS)*100))*100);" over"
369 NEXT X
370 N(1)=1:ERASE N:GOTO 247
600 STOP
1260 INPUT"Enter observation interval length (in secs.)";O:INPUT"Enter time spen
t recording even if zero";O1:REM concurrent categories only
1270 NUMOBS=INT(TIMEALL/(O+O1))
1280 FOR X=0 TO (NUMOBS*(O+O1))-(O+O1) STEP O+O1: FOR X1=X TO X+O-1: M(L(X1))=M(
L(X1))+1
1282 NEXT X1
1285 FOR X2=1 TO 9:IF M(X2)>0 THEN N(X2)=N(X2)+1: M(X2)=0
1286 NEXT X2
1287 NEXT X
1290 C1="intervals": GOTO 360
3000 INPUT"start time (HH:MM)";A:TS=(VAL(LEFT$(A,2))*3600)+(VAL(RIGHT$(A,2))*60)
:IF TS<39900! THEN TS=TS-30600: GOTO 3010
3005 TS=TS-46800!
3010 FOR X= TS TO (TS+1800)
3020 N(L(X))=N(L(X))+1:NEXT X: C1="secs.":NUMOBS=1800:RETURN 360
4000 INPUT "start time (HH:MM:SS)";A: TS=((VAL(LEFT$(A,2)))*3600)+((VAL(MID$(A,4
,2)))*60)+VAL(RIGHT$(A,2)):IF TS<39900! THEN TS=TS-30600:GOTO 4008
4005 TS=TS-46800!
4008 INPUT"Session length (mins.)";L:L=L*60
4010 FOR X=TS TO (TS+L)
4020 N(L(X))=N(L(X))+1:NEXT X: C1="secs.":NUMOBS=L:RETURN 360
10000 OPEN "R",1,G1: FOR V= 0 TO 38: FIELD 1, (V*3) AS AX, 3 AS D(V): NEXT V: FI
ELD 1, 117 AS AX, 2 AS D1, 8 AS B2
10050 GET 1,1:T=CVS(B2)
10100 GET 1, U2: FOR V=0 TO 38: EVENT(W) = LEFT$(D(V),1): TIME(W) =CVI(RIGHT$(D(
V),2))+T: T=TIME(W): IF EVENT(W)= "!" THEN PRINT W;" events input": CLOSE: RETUR
N
10150 IF VAL(EVENT(W))=0 THEN GOTO 10200
10170 W=W+1
10200 NEXT V: U2=U2+1: GOTO 10100
```

```
350 FOR X=03-1 TO (NUMOBS*03)-1 STEP 03:  N(L(X))=N(L(X))+1:NEXT X:C1="points"
360 PRINT" behaviour code      %age of ";C1;"      REAL %age    %age error":FOR
X=1 TO 6:  PRINT"   ";X;"                  ";(N(X)/NUMOBS)*100;"         ";:PRINT
TAB(42) T2(X);: REM for 9 codes change 6 to 9
363 IF T2(X)=0 AND N(X)=0 THEN PRINT TAB(56) " never occurred":GOTO 369
364 IF T2(X)=0 OR N(X)=0 THEN PRINT TAB(56) " infinity ":GOTO 369
365 IF T2(X)/((N(X)/NUMOBS)*100)=1 THEN PRINT TAB(56)" equal":GOTO 369
366 IF T2(X)>(N(X)/NUMOBS)*100 THEN PRINT TAB(56) ((T2(X)-((N(X)/NUMOBS)*100))/T
2(X))*100;" under":GOTO 369
367 PRINT TAB(56) ((T2(X)-((N(X)/NUMOBS)*100))/T2(X))*100;" over"
369 NEXT X
```

```
10 REM simulates other recording systems for sessions of specified starttime an
 duration -- called SIMUL
20 DEFSTR A-G: DEFINT L-O: DEFINT U-Z: DEFSNG T: DIM TIME(1000): DIM EVENT(1000
: DIM D(40): U2=1: DIM L(10800):DIM NTOT(1800)
50 INPUT"Name file for analysis of sampling procedures";G:G1="B:"+G
60 GOSUB 10000
100 FOR X=0 TO W: Z=TIME(X)-TIME(0): L(Z)=VAL(EVENT(X)):NEXT X
150 FOR X=0 TO TIME(W)-TIME(0): IF L(X)=0 THEN L(X)=L(X-1)
160 NEXT X
200 ERASE EVENT: ERASE D:REM input ended a.o.k. up to here l(0 to w) is list of
behs. in each sec. from start of session
230 TIMEALL=TIME(W)-TIME(0)
240 FOR X=0 TO TIMEALL-1: N1(L(X))=N1(L(X))+1: NEXT X: FOR X=1 TO 9: T2(X)=(N1(
)/TIMEALL)*100:NEXT X: REM t2(1-9) are real %ages of occurance
245 INPUT"Start time (HH:MM:SS)";A: TS=((VAL(LEFT$(A,2)))*3600)+((VAL(MID$(A,4,
)))*60)+VAL(RIGHT$(A,2)):IF TS<39900! THEN TS=TS-30600:GOTO 247
246 TS=TS-46800!
247 INPUT"Session length (mins.)";L: TIMEALL=L*60
250  N(1)=1:ERASE N:PRINT"select sampling strategy --- enter 1 for momentary ti
e sample"
255 PRINT"                                enter 2 for interval obs'n method":INPUT
M
257 IF M=2 THEN GOTO 1260
280 INPUT"Enter time in secs. between observations";03:REM point sample starts h
ere
290 NUMOBS=INT(TIMEALL/03)
350 FOR X=TS+(03-1) TO TS+(NUMOBS*03)-1 STEP 03:  N(L(X))=N(L(X))+1:NEXT X:C1="p
oints"
360 PRINT" behaviour code      %age of ";C1;"      REAL %age    %age difference"
:FOR X=1 TO 6: PRINT"  ";X;"                    ";(N(X)/NUMOBS)*100;"          ";:PR
INT TAB(42) T2(X);: REM for 9 codes change 6 to 9
363 IF T2(X)=0 AND N(X)=0 THEN PRINT TAB(56) " 100 equal":GOTO 369
364 IF T2(X)=0 OR N(X)=0 THEN PRINT TAB(56) " zero ":GOTO 369
365 IF T2(X)/((N(X)/NUMOBS)*100)=1 THEN PRINT TAB(56)" 100 equal":GOTO 369
366 IF T2(X)>(N(X)/NUMOBS)*100 THEN PRINT TAB(56) 100-((((N(X)/NUMOBS)*100)/T2(X
))*100);" under":GOTO 369
367 PRINT TAB(56) 100-((T2(X)/((N(X)/NUMOBS)*100))*100);" over"
369 NEXT X
370 INPUT"Enter 'y' for more analysis or 'n' to stop";B:ERASE N: IF B="y" GOTO 2
80 ELSE END
600 STOP
1260 PRINT"needs fixing":STOP:INPUT"Enter observation interval length (in secs.)
";O:INPUT"Enter time spent recording even if zero";O1:REM concurrent categories
only
1270 NUMOBS=INT(TIMEALL/(O+O1))
1280 FOR X=TS TO TS+(NUMOBS*(O+O1))-(O+O1) STEP O+O1: FOR X1=X TO X+O-1: M(L(X1)
)=M(L(X1))+1
1282 NEXT X1
1285 FOR X2=1 TO 9:IF M(X2)>0 THEN N(X2)=N(X2)+1: M(X2)=0
1286 NEXT X2
1287 NEXT X
1290 C1="intervals": GOTO 360
10000 OPEN "R",1,G1: FOR V= 0 TO 38: FIELD 1, (V*3) AS AX, 3 AS D(V): NEXT V: FI
ELD 1, 117 AS AX, 2 AS D1, 8 AS B2
10050 GET 1,1:T=CVS(B2)
10100 GET 1, U2: FOR V=0 TO 38: EVENT(W) = LEFT$(D(V),1): TIME(W) =CVI(RIGHT$(D(
V),2))+T: T=TIME(W): IF EVENT(W)= "!" THEN PRINT W;" events input": CLOSE: RETUR
N
10150 IF VAL(EVENT(W))=0 THEN GOTO 10200
10170 W=W+1
10200 NEXT V: U2=U2+1: GOTO 10100
```

```
20 DEFSTR A-G: DEFINT L-O: DEFINT U-Z: DEFSNG T: DIM TIME(1000): DIM EVENT(1000
: DIM D(40): U2=1: DIM L(1080):DIM NTOT(1800):DIM V(6,900):DIM W(6,900):DIM U(9
0)
50 INPUT"Name file for durations and IRTS";G:G1="B:"+G
60 GOSUB 10000
90 Z=0:W=W-1
100 FOR X=Z TO W:L(X)=VAL(EVENT(X)): IF L(X)=L(X+1) THEN GOSUB 2000
110 NEXT X
120 INPUT"code for analysis";Z: FOR X=0 TO W-1: IF L(X)=Z THEN V(L(X),INT((TIME
X+1)-TIME(X))/10))=V(L(X),INT((TIME(X+1)-TIME(X))/10))+1
130 NEXT X
140 FOR X=900 TO 0 STEP-1:IF V(Z,X)<>0 THEN GOTO 160
150 NEXT X:PRINT"Code ";Z;" didn't occur":GOTO 400
160 FOR Y=0 TO X:IF V(Z,Y)<>0 THEN PRINT"duration is ";Y*10;" -";(Y*10)+10;" se
s.";TAB(35);"N =";V(Z,Y):Z1=Z1+V(Z,Y):V(Z,Y)=0
170 NEXT Y:PRINT TAB(35)"Total N =";Z1:Z1=0
200 Y=0:FOR X=0 TO W-1:IF L(X)=Z THEN U(Y)=X:Y=Y+1
205 NEXT X: IF Y<2 THEN PRINT"IRT not calculable -- only 1 occurrence"
210 FOR X=0 TO Y-2:T5=TIME(U(X+1))-TIME(U(X)+1):W(Z,INT(T5/10))=W(Z,INT(T5/10))
1
220 NEXT X
240 FOR X=900 TO 0 STEP-1: IF W(Z,X)<>0 THEN GOTO 260
250 NEXT X
260 FOR Y=0 TO X:IF W(Z,Y)<>0 THEN PRINT"IRT is ";Y*10;" -";(Y*10)+10;" secs.";
AB(35);"N =";W(Z,Y):W(Z,Y)=0
270 NEXT Y
400 GOTO 120
2000 STOP:FOR Y=X+1 TO W: L(Y)=L(Y+1):TIME(Y)=TIME(Y+1):NEXT Y:W=W-1:Z=X: RETUR
 100
10000 OPEN "R",1,G1: FOR V= 0 TO 38: FIELD 1, (V*3) AS AX, 3 AS D(V): NEXT V: F
ELD 1, 117 AS AX, 2 AS D1, 8 AS B2
10050 GET 1,1:T=CVS(B2)
10100 GET 1, U2: FOR V=0 TO 38: EVENT(W) = LEFT$(D(V),1): TIME(W) =CVI(RIGHT$(D
V),2))+T: T=TIME(W): IF EVENT(W)= "!" THEN PRINT W;" events input": CLOSE: RETU
N
10150 IF VAL(EVENT(W))=0 THEN GOTO 10200
10170 W=W+1
10200 NEXT V: U2=U2+1: GOTO 10100
```

Appendix 2. Example of inter-observer agreement matrix displayed by programme KAPPA. This is a print from the VDU screen.